DOCUMENT RESUME

ED 198 180 TM 810 149

AUTHOR Scriven, Michael
TITLE Evaluation Thesaurus. Second Edition.
PUB DATE 80
NOTE 157p.: This work is one of a series to come out in 1980-81.
AVAILABLE FROM Edgepress, Box 69, Pt. Reyes, CA 94956 ($7.95 plus $.85 postage per single copy).

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Definitions: Educational Assessment: *Evaluation: Evaluation Methods: *Thesauri

ABSTRACT

This thesaurus to the evaluation field is not restricted to educational evaluation or to program evaluation, but also refers to product, personnel, and proposal evaluation, as well as to quality control, the grading of work samples, and to all the other areas in which disciplined evaluation is practiced. It contains many suggestions, procedures, comments, criticisms, definitions and distinctions. Criteria for inclusion of an entry were: (1) at least a few participants in workshops or classes requesting it: (2) a short account was possible: (3) the account was found useful: or (4) the author thought it should be included for the edification or More current slang and jargon have been included then is usual, because that is seen as a problem area. The statistics and measurement area is lightly treated because it is believed to be well covered in other works. Many terms from the federal/state contract process appear because evaluation is often funded in this way. References have been kept to a select few. Acronyms (around 100) appear in an appendix. (PL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED198180

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Michael Scriven

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 149

2

EVALUATION
second edition
THESAURUS

MICHAEL
SCRIVEN

EDGEPRESS INVERNESS CALIFORNIA

Copyright © 1980, Michael Scriven

International Standard Book Number 0-918528-08-9
Second edition
First printing September 1980

Library of Congress, Cataloging in Progress,
Number 80-68775

No part of this book may be reproduced by any mechanical, photographic, or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted, or otherwise copied for public or private use, without permission from the publisher.

Edgepress
Box 69, Pt. Reyes
California 94956

Printed in the United States of America

Single copies $6.95 + postage etc. .55¢
Two or more, 20% off price & postage.
California tax for California order only: 6%
(.42¢ per single copy, .34¢ per copy on multiple orders.)
Mastercharge or Visa acceptable.
Above prices good for second edition, all printings.
Third edition projected late 1981.
Free copies of third edition to contributors who identify significant additions and errors.

INTRODUCTION TO THE SECOND EDITION

Evaluation is a new discipline though an old practice. It is not just a science, though there is a point to talking about scientific evaluation by contrast with unsystematic or subjective evaluation. Disciplined evaluation occurs in scholarly book reviews, the Socratic dialogs, social criticism and in the opinions handed down by appellate courts. Its characteristics are the drive for a determination of merit, worth or value; the control of bias; the emphasis on sound logic, factual foundations and comprehensive coverage. That it has become a substantial subject is attested to by the size of this work and of the work to which the entries herein refer. It is a subject in its own right, not to be dissipated in sub-headings under education, health, law-enforcement and so on; one might as well argue that there is no subject of statistics, only agricultural statistics, statistics in biology etc. Nor will it do to classify evaluation under "Social Sciences, Sundry" since evaluation far transcends the social sciences. That the Library of Congress will not recognize the autonomy of the discipline leads to unflattering (evaluative) conclusions about the bureaucracy of bibliography. But scholars who read the dozen journals and the scores of books, in the field, as well as the government (whose needs are more practical) are coming to recognize evaluation as a subject that requires certain skills, some knowledge and specific training.

This small work may serve as a kind of miniature text-cum-reference-guide to the field. It developed from a 1977 pamphlet entitled ***Evaluation Thesaurus***, and the dictionary definition of the term "thesaurus" still applies to this much

larger, more detailed, and massively rewritten work: "a book containing a store of words or information about a particular field or set of concepts" (***Webster III***); "a treasury or storehouse of knowledge" (***Oxford English Dictionary***). We already have a couple of encyclopedias in sub-fields of evaluation (educational evaluation and program evaluation), and many texts provide brief glossaries. But for most consumers, the texts and larger compendia contain more than they want to know or care to purchase—for they are indeed expensive. The glossaries, on the other hand, are too brief. Here then is a smaller and cheaper guide than the encyclopedias, yet one that is more comprehensive than the glossaries and it is not restricted simply to educational evaluation or to program evaluation. It also refers to product and personnel and proposal evaluation, to quality control and the grading of work samples, and to all the other areas in which disciplined evaluation is practiced. It contains many suggestions and procedures, comments and criticisms, as well as definitions and distinctions. Where it functions as a dictionary, it is in the tradition of Samuel Johnson's ***English Dictionary*** rather than the mighty ***OED;*** academic presses would not have approved his definition of oats ("A grain, which in England is generally given to horses, but in Scotland supports the people") but you and I do. Where this serves as a reference to good practice and not just good usage, it is of course briefer than the special texts or encyclopedias, but it may provide a good starting-point for an instructor who wishes to focus on certain topics in considerable detail and to provide tailored readings on those, while ensuring that students have some source for untangling the rest of the complex conceptual net that covers this field. Students who have sat down and read it cover to cover report on the experience as packing a semester's course into two days.

Smaller than the other texts, yes; more judgmental beyond doubt. But also possibly more open to change; we print short runs at Edgepress so that updating doesn't have to compete with protection of inventory. Send in your corrections or suggestions, and receive a free copy of the next edition. The most substantial or numerous suggestions also earn the choice of a handsome book on evaluation from our stock of spares. (At this writing, we have spares of both encyclopedias and twenty other weighty volumes.)

The criteria for inclusion of an entry were (a) at least a few participants in workshops or classes requested it, (b) a short account was possible, (c) the account was found useful, *or*—in a few cases—(d) the author thought it should be included for the edification or amusement of professionals and/or amateurs. There is much more current slang and jargon in here than would usually be recognized by a respectable scholarly publication—but that's exactly what gives people the most trouble. (And besides, though some of the slang is unlovely, some of it embodies the poetry and imaginativeness of a new field far better than more pedestrian and technical prose.) There's not much on the solid statistics and measurement material—because that's very well covered elsewhere. (But there's a little, because participants in some inservice workshops for professionals have no statistical background and find these few definitions helpful.) There's a good deal about the federal/state contract process because that's the way much of evaluation is funded (and because its jargon is especially pervasive and mysterious). Some references are provided—but only a few key ones, because too many just leave the readers' problem of selection unanswered. The scholar will usually find more references in the few given; that was one criterion for selection of them. Acronyms, besides a basic few, are in a supplement, to reduce clutter. The list of entries has benefited from comparison with the ***Encyclopedia of Educational Evaluation*** (eds. Anderson, Ball, Murphy *et al.*, Jossey Bass, 1976); but there are over 120 substantive entries here that are *not* in ***EEE.***

The University of San Francisco, through its support of the Evaluation Institute, deserves first place in a listing of indebtedness. In 1971–72 the U.S. Office of Education (embodied in John Egermeier) was kind enough to support me in developing a training program in what I then called Qualitative Educational Evaluation at the University of California at Berkeley, and there began the glossary from which this work grew. Two contracts with Region IX of HEW, to assist in building staff evaluation capability, have led me from giving workshops there to developing materials which can be more widely distributed, more detailed, and used for later reference more often than seminar notes; this thesaurus is part of those materials. My students and contacts in those courses and workshops, as others at Berkeley,

Nova, USF, and elsewhere, have been a constant source of improvement—still needed—in formulating and covering this exploding and explosive field; and my colleagues and clients too. To all of these, many thanks, most especially to Jane Roth for her work on the original ***Evaluation Thesaurus*** which she co-authored in 1977, and Howard Levine for many valuable suggestions on the first edition. Thanks, too, to Sienna S'Zell and Nola Lewis for handling the complexities of getting this into and out of our Mergenthaler phototypesetters. They are not to blame for our minor efforts to reform punctuation e.g. by usually omitting the commas around "e.g." since it provides its own pause in the flow; and cutting down on the use of single quotes, since the U.S. and British practices are reversed.

This work is one of a series to come out in 1980-81. A companion monograph, ***The Logic of Evaluation***, is complete and should be available in September. ***The Evaluation of Composition Instruction*** will be available in November. ***Product Evaluation*** is typeset and in field readers' hands and should be available by the end of the year. ***Introduction to Evaluation*** is scheduled for January '81. Others are projected on personnel evaluation, qualitative research methodology, apportionment, etc. (A note to Edgepress or the Institute will ensure that a descriptive flyer on each will be sent as they become available.) Where more details on a topic referenced in the thesaurus are provided in the first two of these monographs, the abbreviations ***LE*** (for ***Logic of Evaluation***) and ***PE*** (for ***Product Evaluation***) are used.

Evaluation Institute
University of San Francisco
California 94117

September 1980

Terms are printed in bold type to indicate that they have their own entry; this slightly distracting flag is not waved more than once in any entry.

ACCOUNTABILITY Responsibility for the justification of expenditures or of one's own efforts. Thus program managers and teachers should be, it is often said, accountable for their costs and salaries and time. The term is also used to refer to a movement towards increased expectations, e.g. of *more* detailed justification of expenditure or efforts. Accountability thus requires some kind of cost-effectiveness evaluation; it is not enough that one be able to explain how one spent the money ("fiscal accountability"), but it is also expected that one be able to *justify* this in terms of the achieved results. Teachers have sometimes been held (wholly) accountable for their students' achievement scores, which is of course entirely inappropriate since their contribution to these scores is only one of several (support from parents, from peers, and from the rest of the school environment outside the classroom are the most frequently cited other influences). On the other hand, a teacher *can* appropriately be held accountable for the failure to produce the same kind of learning gains in his or her pupils that other teachers of essentially similar pupils achieve. A common fallacy associated with accountability is to suppose that justice requires the formulation of precise goals and objectives if there is to be any accountability; but in fact one may be held accountable for what one does, within even the most general conception of professional work, e.g. for "teaching social studies in the twelfth grade," where one might be picking a fresh (unprescribed) topic every day, or every hour, in the light of one's best judgment as to what contemporary social events and the class capabilities make appropriate. (Captains of vessels are held accountable for their actions in wholly unforeseen circumstances.) It is true, however, that any *testing* process has to be very carefully selected and applied if educational accountability is to be enforced in an equitable way; this does not mean that the test must be matched to what is taught (because what is taught may have been wrongly chosen), but it does mean that the test must be very carefully justified, e.g. by reference to reasonable expectations as to what should be (or could justifiably have been) covered, given the need and

ability of the students.

ACCREDITATION The award of credentials, in particular the award of membership in one of the regional associations of educational institutions or one of the professional organizations which attempt to maintain certain quality standards for entry. The "accreditation process" is the process whereby these organizations determine eligibility for membership and encourage self-improvement towards achieving and maintaining that status. The accreditation process has two phases; in the first, the institution undertakes a self-study and self-evaluation exercise against its own mission statement. In the second phase the regional accrediting commission sends in a team of people familiar with similar institutions, to examine the self-study and its results, and to look at a very large number of particular features of the institution, using data to be supplied by the institution together with a checklist (***Evaluative Criteria*** is the best known of these, published by The National Society for School Evaluation), which are then pulled together in an informal synthesis process. At the elementary level, schools are typically not visited (although there is one of the handful of regional accrediting commissions that is an exception to this); at the high school level a substantial team visit is involved, and the same is true at the college level. Accrediting of professional schools, particularly law schools and medical schools, is also widespread and done by the relevant professional organizations; it operates in a similar way. Accrediting of schools of education that award credentials, e.g. for teaching in elementary schools, is done by the state; there is also a private organization which evaluates such schools. There are grave problems with the accreditation process as currently practiced, in particular its tendency towards the rejection of innovations simply because they are unfamiliar (naturally this is denied); its use of teams unskilled in evaluation; its disinterest in looking at learning achievements by contrast with process indicators; the inconsistency between its practice and the claim that it accepts the institution's own goals; the brevity of the visits; the institutional veto and middle-of-the-road bias in selecting team members; the lack of concern with costs; and so on (***LE***). See **Institutional Evaluation.**

ACHIEVEMENT vs. APTITUDE (THE APTITUDE/ ACHIEVEMENT DISTINCTION) It's obvious enough that there's a difference between the two; Mozart presumably had more early aptitude for the piano than you or I, even if he'd never been shown one. But statistical testing methodology has always had a hard time over the distinction because statistics isn't subtle enough to cope with the point of the distinction, just as it isn't subtle enough to cope with the distinction between correlation and causation. For no one has achievement who doesn't have aptitude, by definition, so there's a one-way correlation; and it's very hard to show that someone has an aptitude without giving them a test that actually measures (at least embryonic) achievement. Temerarious testing types have thus sometimes been led to deny that there is any *real* distinction, whereas the fact is only that they lack the tools to detect it. Distinctions only have to be conceptually clear, not statistically simple; and the distinction between a capacity (an aptitude) and a manifested performance (achievement) is conceptually perfectly clear. *Empirically*, we may never find good tests of aptitude that aren't mini-achievement tests. (Ref. ***The Aptitude Achievement Distinction***, ed. Green, McGraw-Hill.)

ACTION RESEARCH A little-known sub-field in the social sciences that can be seen as a precursor of evaluation.

ACTORS Social science (and now evaluation) jargon term for those participating in an evaluation, typically evaluator, client and evaluee (if a person or his/her program is being evaluated). May also be used to refer to all active **stakeholders.**

ADMINISTRATOR EVALUATION A species of **personnel evaluation** which illustrates many of the problems of **teacher evaluation** in that there is no demonstrably superior administrative style (e.g. with respect to democratic versus authoritarian leadership), where the criterion of merit is effectiveness, rather than enjoyability. The three main components of administrative evaluation should be: (a) anonymous holistic rating of observed performance as an administrator, with an opportunity to give reasons, by all those "significantly interactive" with the individuals in question. Identifying this group is done by a preliminary

request for a list from the administrator to be evaluated to which is attached the comment that the search will also be instigated from the groups at the other end of the interaction; (b) a study of objective measures of effectiveness, e.g. turnaround time on urgently requested materials, output indicators, staff turnover etc.; and (c) paper-and-pencil or simulation tests of relevant knowledge and skills, in particular of new knowledge and understanding that has become important since the time of the last review. This kind of evaluation can easily be tied to in-service training, so that it is a productive and supportive experience. The usual farce of administrator evaluation via performance or behavioral objectives is not only a prime opportunity for the con artist to exploit, is not only indefensible because of its lack of input from most of the people that have most of the relevant knowledge, it is also highly destructive of creative management because of the lack of rewards for handling "targets of opportunity"—indeed, there are usually *de facto* punishments for trying to introduce them as new objectives. (It also has the other weaknesses of any **goal-based evaluation.**)

Administrators are often nervous about the kind of approach listed as preferable here, because they rightly understand that most of the people with whom they interact have a pretty poor grasp of the administrator's extensive responsibilities and burdens. The questionnaire must of course rather carefully delimit the requested response to **rating** (holistically) the *observed* behaviors, and the rest of the objection is taken care of by the comprehensive nature of the group interviews, supplemented by the objective measures.

ADVOCATE-ADVERSARY EVALUATION (THE ADVERSARY APPROACH) A type of evaluation in which, during the process and/or in the final report, presentations are made by individuals or teams whose goal is to provide the strongest possible case for or against a particular view or evaluation of the program (etc.). There may or may not be an attempt at providing a synthesis, perhaps by means of a judge or a jury or both. The techniques were developed very extensively in the early seventies, from the initial example in which Stake and Denny were the advocate and the adversary (the TCITY evaluation), through Bob Wolf, Murray

Levine, Tom Owens and others. There are still great difficulties in answering the question, "When does this give a better picture and when does it tend to falsify the picture of a program?" The search for justice—where we rely on the adversary approach—is not the same as the search for truth; nevertheless, there are great advantages about stating and attempting to legitimate radically different appraisals, e.g. the competitive element. One of the most interesting reactive phenomena in evaluation was the effect of the original advocate-adversary evaluation; many members of the "audience" were extremely upset by the fact that the highly critical adversary report had been printed as part of the evaluation. They were unable to temper this reaction by recognition of the equal legitimacy accorded to the advocate position. The significance of this phenomenon is partly that it reveals the enormous pressures towards bland evaluation, whether they are explicit or below the surface. In "purely logical" terms, one might think there wasn't much difference between giving two contradictory viewpoints equal status, and giving a merely neutral presentation. But the effect on the audience shows that this is not the case; and indeed, a more practically oriented logic suggests that important information is conveyed by the former method of presentation that is absent from the latter, namely the *range* of (reasonably) defensible interpretations. See also **Relativism, Judicial Model.**

ADVOCATE TEAMS APPROACH (Stufflebeam) Not to be confused with the advocate/adversary approach to evaluation. A procedure for developing in detail the leading options for a decision maker, as a preliminary to an evaluation of them.

AFFECTED POPULATION A program, product etc. impacts the true **consumers** and its own staff. In program evaluation both effects must be considered though they have quite different ethical standings. At one stage, it looked as if the Headstart program could be justified (only) because of its benefits to those it employed.

AFFECTIVE (Bloom) Original sense; pertaining to the domain of affect. Often taken to be the same as the domain of feelings or attitudes. Since these are sometimes confused with beliefs, it should be remembered that affect should also

be distinguished from the **cognitive** and **psychomotor** domains. For example, self-esteem and **locus of control** are often said to be affective variables, but many items or interview questions which are said to measure these actually call for estimates of self-worth and appraisals or judgments of locus of control, which are straight propositional claims and hence cognitive. Errors such as this often spring from the idea that the realm of valuing is not propositional, but merely attitudinal, a typical fallacy of the **value-free** ideology in social science. While some personal values are evident in attitudes and hence may be considered affect, some valuations—whether or not they cause certain attitudes—are scientifically testable assertions. Note the difference between, "I *feel* perfectly capable of managing my own life, selecting an appropriate career and mate, etc." and "I *am* perfectly capable, etc." (Or "I *feel* this program is really valuable for me." vs. "This program *is* really valuable for me.") Claims about feelings are autobiographical and the error sources are lying and lack of self-knowledge. Claims about merit are external world claims and verified or falsified by evaluations. The use of affective measures, beyond the simplest expressions of pleasure, is currently extremely dubious because of (a) these conceptual confusions between affect and cognition, (b) deliberate falsification of responses, (c) unconscious misrepresentation, (d) dubious assumptions made by the interpreter, e.g. that increases in self-esteem are desirable (obviously false beyond a certain (unknown) point), (e) invasion of privacy, (f) lack of even basic validation, (g) high lability of much affect, (h) high stability of other affect. Not long ago, I heard an expert say that the only known-valid measure of affect relates to locus of control and *that* is fixed by the age of two. He may have been optimistic.

ANALYTICAL (evaluation) By contrast with **holistic evaluation,** which might be called macro-evaluation (by analogy with macro-economics), analytical evaluation is micro-evaluation. There are two main varieties: **component evaluation** and **dimensional evaluation.** It is often thought that causal analysis or remedial suggestions are part of analytic (typically formative) evaluation, but they are not in fact part of evaluation at all (*LE*).

ANCHORING (ANCHOR POINTS) **Rating scales**

that use numbers (e.g. 1-6, 1-10) or letters (A–F) should normally provide some translation of the labeled points on the scale, or at least the end-points and mid-point. It is common, in providing these anchors, to confuse **grading** language with **ranking** language, e.g. by defining A–F as "Excellent ... Average ... Poor" which has two absolute and one relative descriptors, hence is useless if most of the evaluands are or may be excellent (or poor). Some, probably most, anchors for letter grades create an asymmetrical distribution of merit, e.g. because the range of performances which D (potentially) describes is narrower than the B range; this invalidates (though *possibly* not seriously) the numerical conversion of letter-grades to grade points (***LE***). It may be a virtue, if conversion is *not* essential. In another but related sense of anchoring, it means cross-calibration of e.g. several reading tests, so as to identify (more or less) equivalent scores.

ANONYMITY The preservation of the anonymity of respondents sometimes requires very great ingenuity. Although even bulletproof systems do not achieve honest responses from everyone in personnel evaluation, because of **secret contract bias**), leaky systems get honesty from almost no-one. The new legal requirements for open files has further endagered this crucial source of evaluation input; but not without adequate ethical basis. The use of a "filter" (a person who removes identifying information, usually the person in charge of the evlauation) is usually essential; a suggestion box, a phone with a recorder on it to which respondents can talk (disguising their voice), checklists that avoid the necessity for (recognizable) handwriting, forms that can be photocopied to avoid watermark identifiers, money instead of stamps or reply-paid envelopes (which can be invisibly coded), are all possibilities. Typical further problems: What if you want to provide an *incentive* for responding—how can you tell who to reward?; What if, like a vasectomy, you wish to reverse the anonymyzing process (e.g. to get help to a respondent in great distress)? There are complex answers, and the questions illustrate the extent to which this issue in evaluation design takes us beyond standard survey techniques.

APPLES & ORANGES ("Comparing apples & oranges") Certain evaluation problems evoke the complaint, particu-

larly from individuals trained in the traditional social sciences, that any solution would be "like comparing apples and oranges." Careful study shows that *any* true evaluation problem (as opposed to a unidimensional measurement problem) involves the comparison of unlike quantities, with the intent of achieving a **synthesis.** It is the nature of the beast. On the other hand, far from being impossible, the simile itself suggests the solution; we do of course compare apples and oranges in the market, selecting the one or the other on the basis of various considerations, such as cost, quality relative to the appropriate standards for each fruit, nutritional value, and the preferences of those for whom we are purchasing. Indeed, we commonly consider two or more of these factors and rationally amalgamate the results into an appropriate purchase. While there are occasions on which the considerations just mentioned do not point to a single winner, and the choice may be made arbitrarily, this is typically not the case. Complaining about the apples and oranges difficulty is a pretty good sign that the complainer has not thought very hard about the nature of evaluation (*LE*).

APPORTIONMENT (ALLOCATION, DISTRIBUTION) The process or result of dividing a given quantity of resources between a set of competing demands, e.g. dividing a budget between programs. This is in fact the defining problem of the science of economics, but one that is usually not addressed directly or not in practical terms within the economic literature, presumably because any solution requires making assumptions about the so-called "interpersonal comparison of utility," i.e., the relative worth of providing goods to different individuals. Thus the value-free conception of the social sciences makes it taboo to provide practical solutions to the apportionment problem. Apportionment is a separate evaluation predicate, distinct from grading and ranking and scoring although all of those are involved in it; it is one, very practical, way of showing one's estimate of relative worth, and of all the evaluation predicates it is probably the closest to the decision makers' modal evaluation process. Various patently inappropriate solutions are quite frequently used, e.g. the "across-the-board cut." This not only rewards the padding of budgets, and hence automatically leads to increased padding the follow-

ing year, but it also results in some funding at below the "critical mass" level, a complete waste of money. Another inappropriate solution involves asking program managers to make certain levels of cut; this of course results in the blackmail strategy of setting the critical mass levels too high, in order to get more than is absolutely necessary. The only appropriate kind of solution involves some evaluation by a person external to the program, typically in conjunction with the program manager; and the first task of such a review must be to eliminate anything that looks like fat in the budget. Later steps in the process involve segmentation of each program, identification of alternative articulations of the segments, grading of the cost-effectiveness of the progressively larger systems in each sequence of add-ons, and consideration of interactions between program components that may reduce the cost of each at certain points. Given an estimate of the "return value" of the money (the good it would do if not used for this set of programs), and the ethical (or democratic) commitment to *prima facie* equality of interpersonal worth, one then has an effective algorithm for spending the available budget in the most effective way. It will typically be the case that some funding of each of the programs will occur (unless the critical mass is too large), because of the declining marginal utility of the services to each of the (semi-overlapping) impacted populations, the long-term advisability of retaining capability in each area, and the political considerations involved in reaching larger numbers. The process just described provides a rationale for what has sometimes been called zero-based budgeting, an innovation of which the Carter Administration made a good deal in the first years of his presidency; but serious discussion of the methodology for it never seemed to emerge, and the practice was naturally well behind that. At the informal but highly practical level, apportionment reminds us of one of the most brilliant examples of **bias control** methodology in all evaluation; the solution to the problem of dividing an irregularly-shaped portion of food or land into two fair shares—You divide, and I'll choose. (This is a micro-version of the "veil of ignorance" or antecedent probability approach to the justification of justice and ethics in Rawls, ***A Theory of Justice*** (1971), and Scriven , ***Primary Philosophy,*** McGraw-Hill, 1966. It is not surprising that ethics and evaluation share a common border here,

since justice is often analyzed as a distributional concept. (See *LE*)

ARCHITECTURAL EVALUATION Like the evaluation of detective stories and many novels (see **literary criticism**), this field involves a framework of logic and a skin of aesthetics; it is frequently treated as if only one of these components is important. The solution to the problems of traffic flow, the use of durable fixtures that are not overpriced, the provision of adequate floor-space and storage, meeting the requirements of expansion, budget, safety and the law; these are the logical constraints. The aesthetic are no less important and no easier to achieve. Unfortunately architecture has a poor record of learning by experience i.e. poor evaluation commitment; every new school building incorporates errors of the simplest kind, (e.g. classroom entries at the front of the room) and colleges of architecture when designed by their faculty not only make these errors but are often and widely thought to be the ugliest buildings on the campus. (Cf. Evaluators who write reports readable only by evaluators). It is significant that the Ford Foundation's brilliant conception of a center for school architecture has, after several years operation, sunk without a trace.

ARCHIVES Repository of records in which e.g. minutes of key meetings, old budgets, prior evaluations and other **found data** are located.

ARTEFACT (or ARTIFACT) (of an experiment, evaluation, analytical or statistical procedure) An artificial result, one merely due to (created by) the investigatory or analytic procedures used in an experiment, an evaluation, or a statistical analysis, and not a real property of the phenomenon investigated. Typically uncovered—and in good designs guarded against—by using multiple independent methods of investigation/analysis.

ASSESSMENT Often used as a synonym for **evaluation,** but sometimes used to refer to a process that is more focussed on quantitative and/or testing approaches; the quantity may be money (as in real estate assessment), or numbers and scores (as in National Assessment of Educational Progress). People sometimes suggest that assessment is less of a judgmental and more of a measurement process

than certain other kinds of evaluation; but it might be argued that it is simply a case of evaluation in which the judgment is built into the numerical results. Raw scores on a test of no known content or construct validity would not be assessment; it is only when the test is (supposedly) of basic mathematical competence, for example, that reporting the results constitutes assessment in the appropriate sense, and of course the judgment of validity is the key evaluative component in this.

ATTENUATION (Stat.) In the technical sense this refers to the reduction in correlation due to errors of measurement.

ATTITUDES The compound of cognitive and affective variables describing a person's mental set towards another person, thing, or state. It may be evaluative or simply preferential; that is, someone may think that running is good for you, or simply enjoy it, or both; enjoying it does not entail thinking it is meritorious, nor *vice versa*, contrary to many suggested analyses of attitudes. Attitudes are inferred from behavior, including speech behavior, and inner states. No one, including the person whose attitudes we are trying to determine, is in an infallible position with respect to the inference to an attitudinal conclusion, even though that person is in a *nearly* infallible position with respect to his or her own *inner states*. Notice that there is no sharp line between attitudes and cognition; many attitudes are evinced through beliefs (which may be true or false), and attitudes can sometimes be evaluated as right or wrong, or good or bad, in an objective way (e.g. attitudes towards "the world owing one a living," work, women (men), etc.). See **Affective**

ATTRITION The loss of subjects in the experimental or control/comparison groups during the period of the study. This is often so large as to destroy the experimental design—60% loss within a year is not uncommon in the schools. Hence all choice of numbers in the groups must be based upon a good estimate of attrition plus a substantial margin for error.

AUDIENCE (in Robert Stake's sense) A group, whether or not they are the **client(s)**, who will or should see and

may use or react to an evaluation. Typically there are many such, and typically an evaluation report or presentation will need careful planning in order to serve the several audiences reasonably well.

AUDIT, AUDITOR Apart from the original sense of this term, which refers to a check on the books of an institution by an independent accountant, the evaluation use of the term refers to a third party evaluation or external evaluation, often of an evaluation. Hence—and this is the standard usage in California—an auditor may be a **meta-evaluator,** typically serving in a formative and summative role. In the more general usage, an auditor may be simply an external evaluator working either for the same client as the primary evaluator or for another client. There are other occasions when the auditor is halfway between the original kind of auditor and an evaluation auditor; for example, the Audit Agency of HEW (now HHS/ED) was originally set up to monitor compliance with fiscal guidelines, but their staff are now frequently looking at the methodology and overall utility of evaluations. The same is true of GAO and OMB "audits."

BALANCE OF POWER A desirable feature of the social environment of an evaluation, summed up in the formula: "The power relation of evaluator, evaluee and client should be as nearly symmetrical as possible." For example, evaluees should have the right to have their reactions to the evaluation appended to it when it goes to the client. Similarly, the client should undertake to be evaluated if the contract identifies someone else as the evaluee. (School administrators who are not being properly evaluated have no right to have teachers critically evaluated.) **Meta-evaluation** and **goal-free evaluation** are both part of the Balance of Power concept. Panels used in evaluation should exhibit a balance of power, not a lack of **bias**. There are both ethical and political/practical reasons for arranging a balance of power.

BASELINE (data or measures) Facts about the condition or performance of subjects prior to **treatment.** The essential result of the pretest part of the pretest-posttest approach. Gathering baseline data is one of the key reasons

for starting an evaluation before a program starts, something that always seems odd to budgetary bureaucrats. See **Preformative.**

BASIC CHECKLIST The 18-point checklist for evaluating products, programs, etc., to be found under **Key Evaluation Checklist.**

BEHAVIORAL OBJECTIVES Specific goals of, e.g. a program, stated in terms which will enable their attainment to be checked by observation or test/measurement. An idea which is variously seen as 1984/Skinner/dehumanizing, etc., or as a minimum requirement for the avoidance of empty verbalisms. Some people now use "measurable objectives" to avoid the miasma associated with the connotations of behaviorism. In general, people are now more tolerant of objectives that are somewhat more abstractly specified, provided that leading verification/falsification conditions can be spelled out. This is because the attempt to spell everything out (and skip the statement of intermediate-level goals) produces 7633 behavioral objectives for reading, which is an incomprehensible mess. Thus educational research has rediscovered the reason for the failure of the precisely analogous move by positivist philosophers of science to eliminate all theoretical terms in favor of observational terms. The only legitimate scientific requirement here is that terms have a *reliable use* and *agreed-upon empirical content* not a *short translation into observational language*—the latter is just one way to the former. Fortunately, scientific training can lead to the reliable (enough) use of theoretical terms, i.e., they can be unpacked into the contextually-relevant measurable indicators upon demand, thereby avoiding the total loss of the main cognitive organizers (above the taxonomical level) and all understanding that would result from the total translation project, even if it were possible. The same conclusion applies to the use of somewhat general goal statements.

BIAS A condition in an evaluation or other design, or in one of its participants, that is likely to produce errors; for example, a sample of the students enrolled in a school is biased against lower economic groups if it is selected from those present on a particular day since absenteeism rates are usually higher amongst lower economic groups. Hence, if

we are investigating an effect that *may* be related to economic class, using such a sample would be faulty design. It is common and incorrect to suppose that (strong) preferences are biases, e.g. someone who holds strong views against the use of busing to achieve desegregation is often said to be biased. (See the glossary of ***Evaluation Standards***, McGraw Hill 1980; where bias is wrongly defined as *"a consistent alignment with one point of view."* This is true *only* where the views are unjustified, i.e., involve or will probably lead to *errors*. It is *not* true if the views are merely controversial; one would scarcely argue that believers in atoms are biased even though the existence of atoms is denied by Christian Scientists. One sometimes needs a judge in a dispute that is *neutral* or *acceptable to both parties;* this should be distinguished from *unbiased*. Being neutral, etc., is often a sign of *error* in a given dispute i.e. a sign of bias. Evaluation panels should usually include trained and knowledgeable people with strong commitments both for and against whatever approach, program, etc., is being evaluated (where such factions exist) and no attempt should be made to select *only* neutral panelists at the usual cost of selecting ignoramuses or cowards and getting superficial, easily dismissed reports. The neutral faction, if equally knowledgeable, should be represented just as any other faction. Selecting a neutral *chair* may be good psychology or politics, but not because s/he is any more likely to be a good judge.

BIAS CONTROL A key part of evaluation design; it is not an attempt to exclude the influence of definite views but of unjustified, e.g., premature or irrelevant views. For example, the use of (some) external evaluators is a part of good bias control, not because it will eliminate the choice of people with definite views about the type of program being evaluated, but because it tends to eliminate people who are likely to favor it for the irrelevant (and hence error-conducive) reasons of ego-involvement or income-preservation (cf. also **Halo Effect**). Usually, however, program managers object to the use of an external evaluator with a known negative view of programs like theirs, which is to confuse bias with preference. Enemies are one of the *best* sources of useful criticism, not that anyone *enjoys* it. Even if it is politically necessary to take account of a manager's opposition to

the use of a negatively-disposed evaluator, it should be done by adding a second evaluator, also knowledgeable, to whom there is no objection, not by finding someone neutral as such, since neutrality is just as likely to be biased; a key point. Other key aspects of bias control involve further separation of the rewards channel from the evaluation reporting, designing or hiring channel, e.g. by never allowing the agency monitor for a program to be the monitor for the evaluation contract on that program, never allowing a program contractor to be responsible for letting the contract to evaluate that program, etc. The ultimate bias of contracted evaluations resides in the fact that the agencies which fund programs fund most or all of their evaluations, hence want favorable ones, a fact of which evaluation contractors are (usually consciously) aware and which does a great deal to explain the vast preponderance of favorable evaluations in a world of rather poor programs. Even GAO, although effectively beyond this influence for most purposes, is not immune enough for Congress to regard them as totally credible, hence—in part—the creation of the CBO (Congressional Budget Office). The possible merits of an evaluation "judiciary," isolated from most pressures by life-time appointment, deserve consideration. Another principle of bias control reminds us of the instability of independence or externality—today's external evaluator is tomorrow's co-author (or spurned contributor). For more details, see "Evaluation Bias and Its Control," in ***Evaluation Studies Review Annual*** (Vol. 1, 1976, ed. G. Glass, Sage). The possibility of neat solutions to bias control design problems is kept alive in the face of the above adversities by remembering the Pie-Slicing Principle: "You slice and I'll select."

BIG SHOPS The "big shops" in evaluation are the five to ten that carry most of the large evaluation contracts; they include Abt Associates, AIR, ETS, RAND, SDC, SRI, etc. (for translations see the acronym appendix). The tradeoffs between the big shops and the small shops run something like this, assuming for the moment that you can afford either: the big shops have enormous resources of every kind, from personnel to computers; they have an ongoing stability that pretty well ensures the job will be done with at least a minimum of competence; and their reputation is important enough to them that they are likely to meet dead-

lines and do other good things of a paper-churning kind like producing nicely bound reports, staying within budget and so on. In all of these respects they are a better bet, often a much better bet, than the small shops. On the other hand, you don't know who you are going to get to work for you in a big shop, because they have to move their project managers around as the press of business ebbs and flows, and as their people move on to other positions; they are rather more hidebound by their own bureaucratic procedures than a small shop; and they are likely to be a good deal more expensive for the same amount of work, because they are carrying a large staff through the intervals between jobs which are inevitable, no matter how well they are run. A small shop is often carrying a proportionally smaller overhead during those times, and may be working out of a more modest establishment, taking some of their payments in the pleasures of independence. It's much easier to get a satisfactory estimate of competence about the large shops than it is about the small shops; but of course what you do learn about the personnel of a small shop is more likely to apply to the people that do your work. There's an essential place for both of them; small shops simply can't manage the big projects competently, although they sometimes try; and the big shops simply can't handle the small contracts. If some more serious evaluation of the quality of the work done was involved in government review panels—and the increasing strength of GAO in meta-evaluation gives some promise of this—then small shops might fit better into the scheme of things, rather as they do in the management consulting field and in the medical specialties. We are buying a lot of mediocre work for our tax dollar at the moment, because the system of rewards and punishments is set up to punish people that don't deliver (or get delivered) a report on time; but not to reward those who produce an outstanding report by comparison with a mediocre one.

BI-MODAL (Stat.) See **Mode.**

BLACK BOX EVALUATION A term, usually employed pejoratively, that refers to **holistic summative** evaluation, in which an overall and frequently brief evaluation is provided, without any suggestions for improvements, etc. Black box evaluation is frequently extremely valuable (e.g. a consumer product evaluation); is frequently far more valid

than any analytical evaluation that could be done within the same time line and for the same budget; and has the great advantage of brevity. But there are many contexts in which it simply will not provide the needed information e.g. where analytical formative evaluation is required. (Note that black box evaluation may even be extremely useful in the **formative** situation.) Cf. **Engineering Model.**

BOILERPLATE Stock paragraphs or sections that are dumped into RFPs or reports (e.g. from storage in a word-processer) to fill them out or fulfill legal requirements. RFPs from some agencies are 90 percent boilerplate—one can scarcely find the specific material in them.

BUDGET Regardless of the form which particular agencies prefer, it's desirable to develop a procedure for project budgeting that remains constant across projects so that your own staff can become familiar with the categories. It can always be converted into a particular required format if it is thoroughly understood. The main categories might be direct labor costs, other direct costs (materials, supplies, etc.), indirect expenses (space and energy costs), other indirect costs (administrative expenses or "general and administrative" expenses (G&A)). The difference between ordinary overhead and G&A is not sharp, but the idea is that ordinary overhead should be those costs that are incurred at a rate proportional to staff salaries on the project, this proportion being the *overhead rate*, e.g. retirement, insurance, etc. G&A will include indirect costs not directly related to project or staff size (for example, license fees and profit). A number of indirect costs such as accounting services, interest charges, etc., could be justifiably put under either category. See **Costs**.

CAI Computer Assisted Instruction. Computer presents the material or at least the tests on it. Cf. **CMI.**

CALIBRATION Conventionally refers to the process of matching the readings of an instrument against a prior standard. In evaluation would include identification of the correct **cutting scores** (which define the grades) on a new version of a test, traditionally done by administering the old and the new test to the same group of students (half getting

the old one first, half the new). A less common but equally important use is with respect to the standardization of *judges* who are on e.g. a site-visit or proposal-reviewing panel. They should *always* be run through two or three calibration examples, specially constructed to illustrate (a) a wide *range* of merit, (b) common difficulties e.g. (in proposal evaluation) comparing low probability of a big pay-off with high probability of a modest pay-off. While it is not crucial to get everyone to give the same rating (interjudge reliability), indeed it decreases validity, it is highly desirable to avoid: (a) intra-judge inconsistency; (b) extreme compression of an individual's ratings, e.g. at the top, bottom or middle, unless the implications and alternatives are thoroughly understood; (c) drift of each judge's standards as they "learn on the job" (let them sort out their standards on the calibration examples); (d) the intrusion of the panel's possibly turbulent group dynamics into the first few ratings (let it stabilize during the calibration period). While the time-cost of calibration may appear to be serious, in fact it is not, if the development of suitable scales and anchor points is undertaken when doing the calibration examples, since the use of these (plus e.g. **salience scoring**) *greatly* increases speed. And, if anyone really cares about validity, or interpanel reliability (i.e. justice), calibration is an essential step. See also **Anchoring.**

CASE-STUDY METHOD The case-study method is at the opposite end of the spectrum of methods from the survey method. Both may involve intensive or casual testing and/or interviewing; observing, on the other hand, is more characteristic of case study method than of large-scale surveys. The case study approach is typical of the clinician, as opposed to the pollster; it is nearer to the historian and anthropologist than it is to the demographer. Causation is usually determined in case studies by the **modus operandi method,** rather than by comparison of an experimental with a control group, although one could in principle do a comparison case study of a matched case. The case study approach is frequently used as an excuse for substituting rich detail for evaluative conclusions, a risk inherent in **responsive evaluation, transactional evaluation** and **illuminative evaluation.** At its best, a case study can uncover causation where no statistical analysis could; and can block or suggest

interpretations that are far deeper than survey data can reveal. On the other hand, the patterns that emerge from properly done large-scale quantitative research cannot be detected in case studies, and the two are thus naturally complementary processes for a complete investigation of e.g. the health or law enforcement services in a city. See also **Naturalistic.**

CAUSATION The relation between mosquitos and mosquito bites. Easily understood by both parties but never satisfactorily defined by philosophers (or scientists).

CEILING EFFECT The result of scoring near the top of a scale—which makes it harder (even impossible) to improve as easily as from a point further down. Sometimes described as "lack of headroom." Scales on which raters score almost everyone near the top will consequently provide little opportunity for anyone to distinguish themselves by outstanding (comparative) performance. In the language of the stock market, they (the scales plus the raters) provide "all downside risk." (Typical of teacher evaluation forms). Usually they should be reconstructed to avoid this; but not if they correctly represent the relevant range of the rated variable, since then the "upside" differences would simply be a measurement artefact. After all, if all the students get all the answers right, there shouldn't be any headroom above their grades on your scale. (You *might* want to use a different *test*, however, if your task was to get a ranking.)

CENTRAL TENDENCY (Stat.) The misleading technical term for the middle or average of a distribution, as opposed to the extent to which it is spread thin, or lumped, the latter being the **dispersion** or variability of the distribution.

CERTIFICATION A term like credentialing, which refers to the award of some official recognition of status, typically based on a serious or trivial evaluation process. **Accreditation** is another cognomen. The certification of evaluators has recently been discussed rather extensively, and raises a number of the usual problems: who is going to be the super evaluator(s) who decide(s) on the rules of the game (or who lost), what would be the enforcement procedures, how would the cost be handled, etc. Certification is a

two-faced process which is sometimes represented as a consumer-protection device—which it can be—and sometimes as a turf-protection device for the guild members, i.e. a restraint of trade process, which it frequently is. Medical certification was responsible for driving out the midwives, probably at a substantial cost to the consumer; on the other hand, it was also responsible for keeping a large number of complete charlatans from exploiting the public. It certainly contributed to the indefensible magnitude of physicians' and lawyers' salaries/fees; and in this respect is consumer-exploitative. The abuses of the big-league auditors, to take another example, are well-documented in ***Unaccountable Accounting*** by Abraham Briloff (1973). When the state gets into the act, as it does with the certification of psychologists in many states, and of teachers in most, various political abuses are added to the above. In areas such as architecture, where non-certificated and certificated designers of domestic structures compete against each other, one can see some advantages to both approaches; but there is very little evidence supporting a single overall conclusion as to the direction which is best for the citizenry, or even for the whole group of practitioners. A well set up certification approach would undoubtedly be the best; the catch is always in the political compromises involved in setting it up; in other countries, the process is sometimes handled better and sometimes worse, depending upon variations in the political process.

CERTIFICATION (of evaluators) See **Evaluation Registry**

CHECKLIST APPROACH (to evaluation) A checklist identifies all significant relevant dimensions of value, ideally in measurable terms, and may also provide for weighting them according to importance. The checklist provides an extremely versatile instrument for determining the quality of all kinds of educational activities and products. The checklist approach reduces the probability of omitting a crucial factor. It reduces artificial overweighting of certain factors by careful definition of the checklist items, so as to avoid overlap. It also provides a guideline for investigating the thoroughness of **implementation** procedures and it reduces possible **halo effect** and **Rorschach** effect. It does not

require a theory and should avoid depending on one as much as possible. Checkpoints—if there are many—should be grouped under categories that have commonsense or obvious meaning, to facilitate interpretation. A checklist does not usually embody the appropriate combinatorial procedure for cases where the dimensions are highly interactive i.e. where the linear or weighted-sum approach fails: such cases are rare.

CIPP An evaluation model expounded in ***Evaluation and Decision-Making*** by Guba, Stufflebeam et al.; the acronym refers to **C**ontext, **I**nput, **P**rocess and **P**roduct evaluation, the four phases of evaluation they distinguish; it should be noted that these terms are used in a slightly special way. Possibly the most elaborate and carefully thought out model extant; it underemphasized evaluation for accountability or for scientific interest.

CITATION INDEX The number of times that a publication or person is referenced in other publications. If used for personnel evaluation, this is an example of a spurious quantitative measure of merit since e.g. it depends on the size of the field, discriminates against the young, against those working on unfashionable topics, does not in fact identify a third of the Nobel laureates etc. Only possible use is in evaluating the *significance* of a particular publication within a field i.e. in history of ideas research; significance is very loosely related to merit.

CLIENT The person (or agency, etc.) for whom an evaluation is formally done. Usually to be distinguished from **audience** and **consumer.** In social program evaluation "client" may be used to mean "consumer," i.e., the client of the *program* rather than the evaluator; it is better to try to use the term "clientele" for that purpose.

CLIENTELE The population directly served by a program.

CLINICAL PERFORMANCE EVALUATION In the health field, and to an increasing extent elsewhere (e.g. teaching evaluation), the term "clinical" is being used to stress a kind of "hands-on" situation which is typically not well tested by anything like paper and pencil tests. However, it can be very well tested by appropriate **simulations,**

as we have seen in some of the medical Boards exams. It can also be very well tested by carefully done structured observations by trained and **calibrated** observers. If one thinks of a paper and pencil test as a limiting case of a simulation, one realizes the enormous extent to which it depends upon imagination and role-playing skills that few of us possess, in order to be realistic. When one turns to look at standard simulations, one finds that these have inherited a great deal of the artificiality of the paper and pencil tests. For example, they rarely involve "parallel processing," that is, the necessity of handling two or three tasks simultaneously. A serious clinical simulation would start the candidate on one problem, providing charts and histories, and then—just as this was beginning to make sense—a new problem with emergency overtones would be thrust at them, and just before they reached the point of making a preliminary emergency decision on that, a third and even more pressing problem would be thrown at them. Given that there is some anxiety associated with test-taking for most people, one could probably come close to simulating clinical settings in this respect. We have long since developed simulations which involve the provision of supplementary information when requested by the testee, part of the scoring being tied to the making of appropriate requests. But very few signs of careful **job analysis** show up in more advanced simulations where a true clinical performance is of interest.

CMI Computer Managed Instruction. Records are kept by the computer, usually on every test item and every student's performance to date. Important for large-scale individualized instruction. Computer may do diagnosis on basis of test results and instruct student as to materials that should be used next. Extent of feedback to *student* varies considerably; main aim is feedback to course manager(s).

COGNITIVE The domain of the propositionally knowable; consisting of "knowledge-that," or "knowledge-how" to perform *intellectual* tasks.

COHORT A term used to designate one group among many in a study, e.g. "the first cohort" may be the first group to have been through the training program being evaluated. Cf. **Echelon**.

COMPETENCY-BASED An approach to teaching or training which focuses on identifying the competencies needed by the trainee, and on teaching to mastery level on these, rather than teaching allegedly relevant academic subjects to various subjectively determined achievement levels. Nice idea, but most attempts at it either fail to specify the mastery level in clearly identifiable terms or fail to show why that level should be regarded as the mastery level. ("Performance-based" is a cognomen.) C-B Teacher Education (CBTE) was a big deal in mid-70s but the catch was that no one could validate the competencies since **style research** has come up with so little. There is always the subject-matter competency requirement, of course, usually ignored in K-12 teacher training and treated as the only one in the post-secondary domain; but CBTE was talking about *pedagogical* competencies—teaching method skills. See also **Minimum Competency, Mastery.**

COMPLIANCE (check). An aspect of **monitoring.**

COMPONENT (evaluation) A component of an **evaluand** is typically a physically discrete part of it, but more precisely any segment that can be said to relate to others in order to make up the whole evaluand. (Typically, we distinguish between the components and their relationships in talking about the evaluand as a system made up of parts or components.) The **holistic evaluation** of something does not imply any evaluation of its components; and an evaluation of components does not automatically imply an evaluation of the whole evaluand—excellent components for an amplifier will not make a good amplifier unless they are correctly related by design and assembly relationships. But since components are frequently of variable equality, and since we are frequently looking for diagnoses that will lead to improvement, evaluating the components may be a very useful approach to **formative evaluation.** If we can also evaluate the relationships, we may have a very helpful kind of formative evaluation—how helpful will depend upon how self-evident or easily determined the "fixes" for defective components are. Component evaluation is distinguished from **dimensional evaluation,** another kind of **analytical evaluation,** by the relatively greater likelihood of manipulability, in a constructive way, of components by

comparison with dimensions (which may be **statistical artefacts**).

CONCEPTUAL SCHEME A set of concepts in terms of whether one can organize the data/results/observations/evaluations in an area of investigation. Unlike **theories**, conceptual schemes involve no assertions or generalizations (other than the minute presuppositions of referential constancy), but they do generate hypotheses and descriptive simplicity.

CONCLUSION-ORIENTED RESEARCH Contrasted with decision-oriented. Cronbach and Suppes' distinction, between two types of educational research, sometimes thought to illuminate the difference between evaluation research (supposedly decision-oriented) and academic social science research (conclusion-oriented). This view is based on the fallacy of supposing that conclusions about merit and value aren't conclusions, a holdover from the positivist, value-free doctrine that value-judgments are not testable propositions, hence unscientific: and on the fallacy of supposing that all evaluation relates to some decision (the evaluation of many historical phenomena e.g. a reign or a policy does not.)

CONCURRENT VALIDITY The validity of an instrument which is supposed to inform us about the *simultaneous* state of another system or variable. Cf. **predictive validity, construct validity.**

CONFIDENTIALITY One of the requirements that surfaces under the legitimate process considerations in the **Key Evaluation Checklist.** Confidentiality, as it is presently construed, relates to the protection of data about individuals from casual perusal by other individuals, not to the protection of evaluative judgments on an individual from inspection by that individual. The requirement that individuals be able to inspect an evaluative judgment made about them, or at least summaries of these with some attempt at preserving anonymity of the evaluator, is a relatively recent constraint on personnel evaluation. It is widely thought to have undermined the process quite seriously, since people can no longer say what they think of the candidate if they have any worry about the possibility of the candidate inferring their authorship and taking reprisals or

thinking badly of them, if the evaluation was critical. It should be noted that most large systems of personnel evaluation have long since failed because people were unwilling to do this even when complete anonymity was guaranteed. (This was characteristic of the armed services systems.) There is no doubt that amongst universities of the first rank there has been a negative effect; but this mostly shows a failure of ingenuity on the part of personnel evaluation, since there are several ways to preserve complete anonymity, under even the weakest laws, namely those which only blank out the name and title of the evaluator. See also **Anonymity**.

CONFLICT OF INTEREST (COI) One of many sources of bias. An evaluator evaluating his/her own products is involved in a conflict of interest—but the result may still be better than the evaluation done by an external evaluator since the latter's loss of intimate knowledge of and experience with the product may not compensate for lack of ego-involvement. That is, although conflict of interest always hurts *credibility*, it does not always affect *validity*. But since it *may* easily affect validity, it is normally better to use at least a mixture of **internal** and **external** evaluation. In choosing panels for evaluation, the effort to pick panelists who have no conflict of interest is usually misplaced or excessive; it is better to choose a panel with a mix (not even an exact balance) of conflicting interests, since they are likely to know more about the area than those with no interests in it or against it. Financial, personal and social ties are no different from intellectual commitment with respect to COI; all can produce better insights as well as worse judgments. The key to managing COI is requiring that the arguments be public and that their validity be scrutinized and voted on by those with other or no relevant COI. See **Bias.**

CONNOISSEURSHIP MODEL Elliott Eisner's non-traditional method of evaluation is based on the premise that artistic and humanistic considerations are more important in evaluation than scientific ones. No quantitative analysis is used but instead the connoisseur-evaluator observes firsthand the program or product being evaluated. The final report is a detailed descriptive narrative about the subject of the evaluation. Cf. **Literary criticism, Naturalistic, Responsive** and **Models.**

CONSONANCE/DISSONANCE The phenomena of cognitive consonance and dissonance, often associated with the work of the social scientist Leon Festinger, are a major and usually underrated threat to the validity of client satisfaction surveys and follow-up interviews as guides to program or product merit. (The limiting case is the tendency to accept Presidential decisions.) Cognitive consonance, not unrelated to the older notion of rationalization, occurs when the subject's perception of the merit of X is changed by his or her having made a strong commitment to X, e.g. by purchasing it, spending time taking it as therapy, etc. Thus a Ford Pinto may be rated as considerably better than a VW Rabbit *after* it has been purchased than before, although no new evidence has emerged which justifies this evaluation shift. This is the conflict of interest side of the coin whose other side is increased knowledge of the (e.g.) product. Some approaches to discounting this phenomenon include very careful separation of needs assessment from performance assessment, the selection of subjects having experience with both (or several) options, serious task-analysis by the *same* trained observers, looking at recent purchasers of both cars, etc. The approval of boot camp by Marines and of cruel initiation rites by fraternity brothers is a striking and important case—called "initiation-justification" bias in ***LE.*** (These phenomena also apply at the meta-level, yielding spurious positive evaluations of evaluations by clients.)

CONSTRUCT VALIDITY The validity of an instrument (e.g. a test, or an observer) as an indicator of the presence of (a particular amount of) a theoretical construct. The construct validity of a thermometer as an indicator of temperature is high, if it has been correctly calibrated. The key feature of construct validity is that there can be no simple test of it, since there is no simple test of the presence or absence of a theoretical construct. We can only infer to that presence from the interrelationships between a number of indicators and a theory which has been indirectly confirmed. The contrast is with predictive and concurrent validity, which relate the readings on an instrument to another directly observable variable. Thus, the predictive validity of a test for successful graduation from a college, administered before admission, is visible on graduation day some years later. But the use of a thermometer to test temperature

cannot be confirmed by looking at the temperature; in fact, the thermometer is as near as we ever get to the temperature. Over the history of thermodynamics, we have adopted four successive different theoretical definitions of temperature, although you couldn't tell this from looking at thermometers. Thus, what the thermometer has "read" has been four different theoretical constructs and its validity as an indicator of one of these is not at all the same as its validity as an indicator of another. No thermometer reads anything at all in the region immediately above absolute zero, since all gases and liquids have solidified by that point; nevertheless, this is a temperature range; and we infer what the temperature is, there, by complicated theoretical calculations from other variables. The validity of almost all tests used for evaluative purposes is construct validity, because the construct towards which they point (e.g. "excellent computational skills") is a complex construct and not observable in itself. This follows from the very nature of evaluation as involving a synthesis of several performance scales. But of course it does not follow that evaluative conclusions are essentially less reliable than those from tests with demonstrated predictive validity, since predictive validities are entirely dependent upon the persistence through time (often long periods of time) of a relationship—a dependency which is often shakier than the inference to an intellectual skill such as computational excellence from a series of observations of a very talented student faced with an array of previously unseen computational tasks. Thermometers are highly accurate though they "only" have construct validity. Construct validity is rather more easily attainable with respect to constructs which figure in a **conceptual scheme** that does *not* involve a theory; only the requirements of taxonomical merit (clarity, comprehensiveness, insight, fertility etc.) need to be met, not confirmation of the axioms and laws of the theory. (Such constructs are still called "theoretical constructs," perhaps because conceptual schemes shade and evolve into theories so fluidly.)

CONSULTANT Consultants are not simply people hired for advice on a short-term basis, as one might suppose from the term; they include a number of people who are essentially regular (but not tenured) staff members of state agencies, where some budgetary or bureaucratic restriction

prevents the addition of permanent staff, but allows a semi-permanent status to the consultant. Hence an evaluation consultant is *not* always an external evaluator. The basic problem about being an evaluation consultant, as a career, is that—with the exception of the semi-permanent jobs just mentioned—you have to make enough on the days you're working to carry you through the days when you're not, and in the real world it is highly unlikely that jobs will be kind enough to fill your time exactly. Meanwhile, some of your overhead, e.g. secretarial and rent, will continue, as well as your grocery bills, etc. Consequently, the most cost-effective consultants from the client's point of view tend to be people with full-time jobs who do their consulting as moonlighting. In the management consultant field, where fees are very much higher than in the evaluation consultant field—almost as high as a regular attorney's fees—this is less of a problem; but in the human services program evaluation area, the true cost of the best consultant is usually far beyond the budgetary limits placed on consulting fees by agencies. It is high time that some system of payment by results was allowed as an alternative, so that there would be some incentive for fast and extremely good work by full-timers, instead of spreading the work out and moonlighting it. The big shops have some full-time evaluators on staff, but only for big projects funded by agencies, not as consultants for the average small client.

CONSUMER The "true consumers" are the persons who are being directly *or* indirectly affected at the using or receiving end of a product or program—the *impacted* populations. The true consumers are not usually just the *target* population. The "consumers" of an evaluation are its **audiences.** The staff of a program are also *affected* by the program, but at the producing or providing end.

CONSUMER-BASED EVALUATION An approach to the evaluation of (typically) a program, that *starts with* and *focuses on* the impact on the consumer or clientele or—to be more exact—the impacted population. It might or might not be done **goal-free**, though clearly that is the methodology of choice for consumer-based evaluation. It will particularly focus on the identification of non-target populations that are impacted, on unintended effects, on true cost to the consumer etc.

CONTENT ANALYSIS The evaluative or pre-evaluative process of systematically determining the characteristics of a body of material or practices, e.g. tests, books, courses, jobs. A great many techniques have been developed for doing this, running from frequency counts on words of certain kinds (e.g. personal references), to analysis of plot structure in illustrative stories to determine whether the dominant figure is e.g. male or female, white or non-white. The use of content analysis is just as important in determining whether the evaluand matches the "official" description of it, as it is in determining what it is and what it does in other dimensions than those involved in the "truth in packaging" issue. Thus, a social studies chart entitled "Great Americans" could be subject to content analysis in order to determine whether those listed were actually great Americans (truth in labeling); but even if it passed that test, it would be subject to further content analysis for e.g. sexism, because a list that did not contain the names of the great women suffragists would show a deformed sense of values, although it might be too harsh to argue that it was not correctly labeled. Notice that none of this refers to a study of the actual effects (**pay-off evaluation**), but is a type of legitimate **process evaluation**. The line between the two is not sharp, since literal falsehoods may be the best pedagogical device for getting the student to remember truths. Although this approach would then violate the requirement of scientific or disciplinary integrity (a process consideration), this would be excused on the grounds that the only point of the work is to produce the right effects and that teaching the correct and much more complicated account leads to less accurate residual learning than teaching the incorrect account. It is not an exaggeration to say that most elementary science courses follow the model of teaching untruths in order to get approximate truths instilled in the brains of the students. A more radical view would hold that human brains in general require knowledge to be presented in the form of rather simple untruths rather than true complexities. An excellent brief discussion of content analysis by Sam Ball will be found on pp. 82-84 of the ***Encyclopedia of Educational Evaluation, which he co-edited for Jossey-Bass, 1976.***

CONTENT VALIDITY The property of tests that, after

appropriate **content analysis**, appear to meet all requirements for congruence between claimed and actual content. Thus a test of net-making ability should contain an adequate (weighted) sampling of *all* and *only* those skills which the expert net-maker exhibits. Note that this is an example of a mainly psychomotor domain of skills; content validity is not restricted to the cognitive or verbal areas. Content validity is one step more sophisticated than **face validity** and one step less sophisticated than **construct validity**. So it can be seen as a more scientific approach to face validity or as a less-than-comprehensive approach to construct validity. The kind of evaluation that is involved in and leads to credentialing by the state as a teacher of e.g. mathematics (in the U.S.) is content *invalid* because of its grotesque failure to require mathematical skills at anything like a reasonable level (e.g. same level as the second quartile of college sophomores majoring in mathematics). In general, like other forms of process evaluation, content validity checks are considerably quicker than construct validity approaches, and frequently provide a rather highly reliable *negative* result, thereby avoiding the necessity for the longer investigation. They cannot provide a positive result so easily, since content validity is a necessary but not a sufficient condition for merit.

CONTEXT (of evaluation or evaluand). The ambient circumstances that do or may influence the outcome.

CONTRACT **See Funding.**

CONTRACT TYPES The usual categories of contract types (this particular classification comes from the Eckman Center's ***The Project Manager's Workplan (TPMWP)***) are fixed price, time and materials, cost reimbursement, cost plus fixed fee, cost plus incentive fee, cost plus sliding fee and joint powers of agreement. Explaining the differences beyond those obvious from the terms would be telling you more than you want to know unless you are about to become a large-project manager, in which case you'll need ***TPMWP***, and may be able to afford it (price upward of $30); it can be ordered from The Eckman Center, P.O. Box 621, Woodland Hills, CA 91365. That's the technical stuff; but at the commonsense level, it's a good idea to have something in writing that covers the basics like when payments are to

be made (and under what conditions they will not be made) and who is empowered to release the results (and when). Dan Stufflebeam has the best checklist for this, in his forthcoming (1981) text.

CONTROL GROUP A group which does not receive the "treatment" (e.g. a service or product) being evaluated. (The group which does receive it is the *experimental group*, though the study may be **ex post facto** and not experimental.) It is used to check the extent to which the same effect occurs without the treatment—which would tend to show the treatment was not causing whatever changes were observed in the experimental group. To do this, the control group must be "matched," i.e., so chosen as to be closely similar to the experimental group (not identical, just similar). The more carefully the matching is done (e.g. by using "identical twins"), the more sure one can be that differences in outcome are due to the experimental treatment. A great improvement is achieved if you can *randomly assign* matched subjects to the two groups, and arbitrarily designate one as the experimental and the other as the control group. This is a "true experiment"; other cases are weaker and include ex post facto studies. Matching would ideally cover all environmental variables as well as genetic ones—all variables *except* the experimental one(s)—but in practice we match only on variables which are likely to significantly affect the results significantly, for example, sex, age, schooling. Matching on specific characteristics (**stratifying**) is not essential, it is only efficient: a perfectly good control group can be set up by using a (much larger) random sample of the population as the control group (and also for the experimental or treatment group). The same degree of confidence in the results can thus be achieved either by comparing small closely matched groups (experimental and control) or large entirely randomly selected groups. Of course, if you're likely to be wrong—or if you're in doubt—about which variables to match on, the large random sample is a better bet even though more expensive and slower. It should be noted that it is sometimes important to run several control groups and that one could then just as well call them all experimental groups or comparison groups. The classical control group is the "no treatment" group, but it's not usually the most relevant to practical decision-making

(see **Critical Competitor**). Indeed, it's often not even clear what "no treatment" *means:* e.g. if you withhold *your* treatment from a control group in evaluating psychotherapy, they create their own, and may change behavior just *because* you withheld treatment—they may get divorced, change or lose their job, etc. So you finish up comparing psychotherapy with *something else*, usually a mixture of things, not with *nothing*, not even with no psychotherapy, only with no psychotherapy of your particular brand. Hence it's better to have control groups that get one or several *standard* alternative treatments than "leave them to their own devices" into which the "no treatment" group often degenerates. And in evaluation, that's exactly where you bring in the critical competitors. In medicine, that's why the control group gets a placebo.

CONVERGENCE GROUP (Stufflebeam). A team whose task is to develop the best version of a treatment from various **stakeholder** or **advocate** suggestions. A generalization of the term, to convergence *sessions*, covers the process that should follow the use of parallel (teams of) evaluators, viz. the comparison of their *written* reports and an attempt to resolve disagreements. This should be done in the first place by the separate teams, with a referee (group) present to prevent bullying; it may later be best to use a separate convergence (synthesis) group.

CORRECTION FOR GUESSING In multiple-choice exams with n alternatives in each question, the average testee would get 1/n of the marks by guessing alone. Thus if a student fails to complete such an exam, it has been suggested that one should add 1/nth of the number of unanswered questions to his or her score, in order to get a fair comparison with the score of a testee that answers all the questions by guessing the ones they do not have time to do seriously. There are difficulties both with this suggestion ("applying the correction for guessing") and with not using it; the correct procedure will depend on a careful analysis of the exact case. Another version of the correction for guessing involves subtracting the number of answers that one would expect to get by guessing from the total score, whether the test is completed or not. These two approaches give essentially the same results, but their effects may interact differently with different instructions on the test: in

general, ethics requires that if such corrections will be used, they be pre-explained to testees.

CORRELATION The relationship of concomitant occurrence or variation. Its relevance to evalution is (a) as a hint that a causal relation exists (showing an **effect** to be present), (b) to establish the validity of an **indicator**. The range is from −1 to +1, with 0 showing random relationship, ±1 showing perfect (100%) correlation (+1) or perfect avoidance (−1).

COSTS, COST-ANALYSIS It is often useful to distinguish initial (start-up) costs from running (maintenance) costs; capital costs from cash flow; discounted from raw costs; direct from indirect costs or overhead, which includes depreciation, maintenance, taxes, some supplies, insurance, some services, repairs, etc.; psychological from tangible costs; outlays from opportunity costs. The "human capital" or "human resources" approach stresses one non-monetary component. "Marginal analysis" looks at the relative *add-on* costs, from a given cost-level, and is often both more relevant to a decision-maker's choices at that basic cost-level, and more easily calculated. Cf. ***Zero-Based Budgeting***.

COST-BENEFIT OR BENEFIT-COST ANALYSIS Cost-benefit analysis goes a step beyond cost-effectiveness analysis (see below) and estimates the overall cost and benefit of each alternative (product or program) *in terms of a single quantity, usually money*. This analysis will provide an answer to the question: Is this program or product worth its cost? Or, which of the options has the highest benefit/cost ratio? (It is often not possible to do cost-benefit analysis, e.g. when ethical, essential, temporal, or aesthetic elements are at stake.)

COST-EFFECTIVENESS ANALYSIS The purpose of this type of analysis is to determine what a program or procedure *costs*, and what it *does* (effectiveness), the latter often being described in terms of qualities (pay-offs) which cannot be reduced to money terms, or to any other single dimension of pay-off. This procedure does not provide an automatic answer to the question: Is this program or product worth its cost? The evaluator will have to weight and synthesize the needs data with cost-effectiveness results to

get an answer, and even that may not give an unequivocal result.

COST-FEASIBILITY ANALYSIS Determining on a Yes/No basis whether something can be afforded (this means you can afford the initial *and* the continuing costs).

COST-FREE EVALUATION The doctrine that evaluations should, if properly designed *and used*, provide a *net positive* return, *on the average*. They may do this by leading either to the elimination of ineffective programs or procedures, or to an increase in productivity or quality from existing resources/levels of effort. The equivalence tables between costs and benefits should be set up to match the client's values, and accepted by the client, before the evaluation begins, so as to avoid undue pressure to be cost-free by cost-cutting only, instead of by quality-improvement as well as cost-cutting (if the latter is requested at all).

COST PLUS Another basis for calculating budgets on contracts is the "cost plus" basis, which allows the contractor to charge for costs plus a margin of profit; depending on how **"profit"** is defined, this may mean the contractor is making less than if the money was in a savings account and s/he was on a salary at some other job, or a good deal more. Sometimes cost plus contracts, since they usually omit any real controls to keep costs down (indeed, sometimes the reverse, since the "plus" is often a *percentage* of the basic cost), are not ideal for the taxpayer either. Which has prompted the introduction of the "cost plus fixed fee" basis, where the fee is fixed and not proportional to the size of the contract. That's sometimes better, but sometimes—when the scope of work is enlarged during the project, by the discovery of difficulties or (subtly) by the agency—it shrinks the profit below a reasonable level. The profit, after all, has to carry the contractor through periods when contracts happen not to abut perfectly, pay the interest on the capital investment, and provide some recompense for high risk. The justification for cost plus contracts is very clear in circumstances where it is difficult to foresee what the costs will be and no sane contractor is going to undertake something with an unknown cost. Especially if the agency wishes to retain the option of changing the conditions that are to be met, the hardware that is to be used, etc., say in the light of

obsolescence of the materials available at the beginning, the cost plus percentage contract can make sense. Competitive bidding is still possible, after all.

CREDIBILITY Evaluations often need to be not only valid but such that their audiences will believe that they are valid (cf. "It is not enough that justice be done, etc."). This may require extra care about avoiding (apparent) **conflict of interest,** for example, even if in a particular case it does not in fact affect validity.

CRITERION The criterion is whatever is to count as the "pay-off," e.g. success in college is often the "criterion measure" against which we validate a predictive test like a college entrance examination. Ability to balance a checkbook might be one "criterion behavior" against which we evaluate a practical math course.

CRITERION-REFERENCED TEST This type of test provides information about the individual's (or a group's) knowledge or performance on a specific criterion. The test scores are thus interpreted by comparison with pre-determined performance criteria rather than by comparison with a reference group (see ***Norm-Referenced Test***). The *merit* of such tests depends completely on the (educational) significance of the criterion—trivial criterion, trivial test; theory-impregnated criterion, theory-dependent test—and on the technical soundness of the test. It is not within an amateur's or the usual teacher's domain of competence to construct such tests, and when they do the results are often uninterpretable because we know *neither* whether the subject understood the question *nor* whether s/he should be able to answer it. It is clear that successful construction of such tests is also beyond the capacity or interest of most professionals: we still lack one good functional literacy test, let alone four or five to choose from.

CRITICAL COMPETITORS Critical competitors are those entities with which comparisons need to be made when a program, product, etc., is being evaluated. The critical competitors can be real or hypothetical, e.g. another existing text or one we could easily make with scissors and paste. They bear on the question whether the *best* use of the money (and other resouces) involved is being made, as

opposed to the pragmatically less interesting question of whether it's just being thrown away. You don't just want to know whether this $20.00 text is *good;* you want to know if there's a much better one for $20.00, or one that is just as good for $10.00. Those others are (two of) the critical competitors that should figure in the evaluation of the text. So should a film (if there is one), lectures, TV, a job or internship, etc., where they or an assemblage of them cover similar material. Traditional evaluation design has tended to use a no-treatment control group for the comparison, which is incorrect; "no treatment" is rarely the real option. It's either the *old* treatment or *another innovative one,* or both, or a *hybrid,* or something no one has so far seen as relevant (or perhaps not even put together). These unrecognized or "created" critical competitors are often the most valuable contributions an evaluator makes and coming up with them requires creativity, local knowledge and realism.

CRITICAL INCIDENT TECHNIQUE (Flanagan) This approach, tied to the analysis of longitudinal records, attempts to identify significant events or times in an individual's life (or an institution's life, etc.) which in some way appear to have altered the whole direction of subsequent events. It offers a way of identifying the effects of e.g. schooling, in circumstances where a full experimental study is impossible. It is, of course, fraught with hazards. (Ref. John Flanagan, ***Psychological Bulletin,*** 1954, pp. 327–358.)

CROSS-SECTIONAL (study) If you want to get the results that a **longitudinal study** would give you, but you can't wait around to do one, then you can use a cross-sectional study as a substitute whose validity will depend upon certain assumptions about the world. In a cross-sectional study, you look at today's first year students and today's graduating seniors and infer e.g. that college has produced the difference between them; in a longitudinal study you would look at today's first year students and wait and see how they change by the time they become graduating seniors. The cross-sectional study substitutes today's graduating seniors for a population which you cannot inspect for another four years, namely the seniors that today's freshman or first year students will become. The assumptions involved are that no significant changes in the demo-

graphics have occurred since the present seniors formed the entering class, and that no significant changes in the college have occurred since that time. (For certain inferences, the assumptions will be in the other direction in time.)

CRYPTO-EVALUATIVE TERM A term which appears to be purely descriptive, but whose meaning necessarily (definitionally) involves evaluative concepts e.g. intelligent, true, deduction. Cf. **Value-imbued.**

CULTURE-FAIR/CULTURE-FREE A culture-free test avoids bias for or against certain cultures. Depending upon how generally culture is defined, and the test is used, this bias may or may not invalidate the test. Certain types of problem-solving tests involving finding food in an artificial desert to avoid starvation, for example, are about as near to culture-free as makes any sense; but they are a little impractical to use. To discover that a test discriminates between e.g. races with respect to the numbers who pass a given standard, has absolutely no relevance to the question of whether the test is culture-fair. If a particular race has been oppressed for a sufficiently long time, then its culture will not provide the kind of support for intellectual exercises (or athletic ones, depending upon the type of oppression); it will probably not provide the dietary prerequisites for full development; and it may not provide the role models that stimulate achievement in that direction. Hence, *quite apart from any affects on the gene pool*, it is to be expected that that racial group will perform worse on certain types of tests—if it did not, the argument that serious oppression has occurred would be weakened. Systematic procedures are now used to avoid clear cases of cultural bias in test items, but these are poorly understood. Even distinguished educators will sometimes point to the occurrence of a term like "chandelier" in a reading vocabulary test as a sign of cultural bias, on the grounds that oppressed groups are not likely to have chandeliers in their houses. Indeed they are not, but that's irrelevant; the question is whether the term reliably indicates wide reading, and hence whether a sufficient number of the oppressor group in fact picked up the term through labeling an object in the environment rather than through wide reading to invalidate that inference. That's an empirical question, not an *a priori* one. A similar point comes up in

looking at the use of test scores for admission selection; validation of a cut-off is properly based on prior experience, and may be based on a mainly white population. In such a case, the use of the same cutting scores for minorities will tend to *favor* them, as a matter of empirical fact (possibly because the later efforts of those individuals get less peer/home support than in the white population.

CURRICULUM EVALUATION Curriculum evaluation can be treated as a kind of *product evaluation*, with the emphasis on outcome studies of those using the curriculum; or it can be approached in terms of **content validity**. ("Curriculum" can refer to the content or to the sequencing of courses, etc.) A popular fallacy in the area involves the supposition that good tests used in a curriculum evaluation should match the goals of the curriculum or at least its content; on the contrary, if they are to be tests *of* the curriculum, they must be independently constructed, by reference to the needs of the user population and the general domain of the curriculum, without regard to its specific content, goals and objectives. Another issue concerns the extent to which long-term effects should be the decisive ones; since they are usually inaccessible because of time or budget considerations, it is often thought that judgments about curricula cannot be made reliably. But essentially all long-term effects are best predicted by short-term effects, which *can* be measured. And the causal inferences involved from temporally remote data, even if we could wait to study the long-term situation, are so much less reliable that any gains from the long-term study would likely be illusory. One of the most serious errors in a great deal of curriculum evaluation involves the assumption that curricula are implemented in much the same way by different teachers, or in different schools; even if a quite thorough checklist is used to ensure implementation, there is still a great deal of slippage in the teaching process. In the more general sense of curriculum, which refers to the sequence of courses taken by a student, the slippage occurs via the granting of exceptions, the use of less-than-valid challenge exams, the substitution of different instructors for others on leave, etc. Nevertheless, good curriculum materials and good curriculum sequences should be evaluated for gross differences in their effectiveness and veracity/comprehensive-

hess/relevance to the needs of the students. The differences between good and bad are so large and common that, despite all the difficulties, very much improved versions and choices can result from even rough and ready evaluation of content and teachability. Davis identifies the following components in curriculum evaluation: determining the actual nature of the curriculum (and its support system of counselors, other curricula, catalogs, etc.) as compared with the official descriptions (e.g. via transcript analysis, curriculum analysis of class notes); evaluating its academic quality; examining procedures for its evaluation and revision; assessing student learning; student surveys including exit of alumni interviews; faculty surveys; surveys of employers and potential employers; reviews by professional curriculum experts; comparison with any standards provided by relevant professional associations; checking with leading schools or colleges to see if they have improvements/updates that should be considered. Ref. ***Designing and Evaluating Higher Education Curriculum,*** Lynn Wood & Barbara Gross Davis, AAHE, 1978.

CUTTING SCORE A score which marks the line between grades, between mastery and non-mastery, etc. Always arbitrary to some degree, it is justifiable in circumstances where a number of such scores will be synthesized eventually. But in a final report, only cutting *zones* make sense and the grades should indicate this, e.g. A, A−, AB, B+, ... where the AB indicates a borderline area. Many opponents of minimum competency testing complain about the arbitrariness of any cut-off *point;* the response should be to use a zone, i.e., three grades (clearly not competent; debatably competent; clearly competent).

DATA SYNTHESIS The semi-algorithmic semi-judgmental process of producing comprehensible facts from raw data via descriptive or inferential statistics and interpretation in terms of concepts, hypotheses or theories.

DECILE (Stat.) See **Percentile.**

DECISION-MAKER It is sometimes important to distinguish between making decisions about the *truth* of various propositions, and making decisions about the *disposi-*

tion of (or *appropriate action* about) something. While the scholar automatically falls into the first category, s/he typically only serves as a consultant to a decision-maker of the second type. Most discussion about decision-makers in the evaluation context refers to those with the power to dispose, not merely with the power to propose or draw conclusions.

DECISION-ORIENTED RESEARCH See ***Conclusion-Oriented Research.***

DECISION RULE A link between an evaluation and action, e.g. "those with a grade below C must repeat the course"; "Hypotheses which are not significant at the .01 level will be abandoned." (The latter example is common but logically improper; see **Null Hypothesis.**)

DELIVERY SYSTEM The link between a product or service and the population that needs or wants it. Important to distinguish this in evaluation, because it helps avoid the fallacy of supposing that the existence of the need justifies the development of something to meet the need. It does so *only if* one can either develop a new (or make use of an existing) delivery system.

DELPHI TECHNIQUE A procedure used in group problem solving, involving—for instance—circulating a preliminary version of the problem to all participants, calling for suggested rephrasings (and/or preliminary solutions). The rephrasings are then circulated for a vote on the version that seems most fruitful (and/or the preliminary solutions are circulated for rank ordering). When the rank orderings have been synthesized, *these* are circulated for another vote. Innumerable variations on this procedure are practiced under the title "Delphi Technique," and there is a considerable literature on it. It is often done in a way that over-constricts the input, hence is ruined before it begins. In any case, the intellect of the organizer must be the equal of the participants or the best suggestions won't be recognized as such. A phone conference call may be more effective, faster and cheaper, perhaps with one chance at written after-thoughts.

DEMOGRAPHICS The characteristics of a population defined in terms of its macroscopic features—age, sex, level

of education, occupation, place of birth, residence, etc., by contrast with micro-features, e.g. IQ, attitude, scores.

DEPENDENT VARIABLE One which represents the outcome—contrast is with the independent variables which are the ones we (or nature) can manipulate directly. That definition is circular and so are all others; the distinction between dependent and independent variables is an ultimate notion in science, definable only in terms of other such notions, e.g. **randomness.**

DESCRIPTIVE STATISTICS The part of statistics concerned with providing illuminating perspectives on or reductions of a mass of data (cf. **inferential statistics**); typically this can be done as a translation, involving no risk. For example, calculating the mean score of a class from its individual scores is straight deduction and no probability is involved. But *estimating* the mean score of the class by calculating the actual mean of a random sample of the class is of course inferential statistics.

DESIGN (of evaluation; see **Evaluation Design**)

DIFFUSION The process of spreading information about (typically) a product (cf. **dissemination** with which diffusion is deliberately and somewhat artificially contrasted).

DIMENSIONAL EVALUATION A species of **analytical evaluation** in which the meritorious performance is broken out into a set of dimensions that have useful statistical properties (e.g. independence) or are familiar from other contexts and easily grasped, etc. Cf. **Component Evaluation.**

DISCREPANCY EVALUATION (Provus) Evaluation conceived of as identifying the gaps between time-tied objectives and actual performance, on the dimensions of the objectives. A slight elaboration of the simple **goal-achievement** model of evaluation.

DISPERSION (Stat.) The extent to which a distribution is "spread" across the range of its variables, as opposed to where it is "centered"—the latter being described by measures of "central tendency," e.g. **mean, median, mode.** Dispersion is measured in terms of e.g. **standard deviation**

or **semi-interquartile difference.**

DISSEMINATION The process of distributing (typically) a product itself, rather than information about it (cf. **diffusion**). Also used as jargon synonym for distribution.

DISSONANCE See **Consonance.**

DOMAIN-REFERENCED TESTING The purpose of testing is not usually to determine the testee's ability to answer the questions on the test, but to provide a basis for conclusions about the testee's ability with regard to a much wider *domain*. Criterion-referenced tests identify ability to perform at a certain (criterion) level on—typically—a particular *dimension*, e.g. two-digit multiplication. DRT is a slight generalization of that to cover cases like social studies education where it seems misleading to suggest that there is *a* criterion. One can think of a domain as defined by a large *set* of criteria, from which we sample, just as—at the other end—the test samples from the testee's abilities. The major problem with DRT is defining domains in a useful way. J.R. Popham has a usefully specific discussion in his ***Educational Evaluation,*** Prentice-Hall, 1975.

DUMPING The practice of unloading funds rapidly near the end of the fiscal year in order that they will not be returned to the central bureaucracy, which would be taken as a sign that next year's budget could be reduced by that amount since it wasn't needed. This may be done with all the trappings of an RFP, i.e., via a contract, but it's a situation where the difference between a contract and a grant tends to evaporate since the contract is so unspecific (because of lack of time for writing the RFP carefully) that it has essentially the status of a grant.

ECHELON A term like "cohort," sometimes used interchangeably with the latter, but better restricted to a group (or group of groups) that is time-staggered with regard to its entry. If a new group comes on board every four weeks for five months, followed by a three month gap, while they are being trained, and then the whole process begins again, the first three groups are called the first echelon; each of them is a cohort.

EDUCATIONAL ROLE (of the evaluator) It is both empirically and normatively the case that this role is of the greatest importance, at worst second only to the truthfinding role. This is not merely because few people have been properly educated as to either the importance or the techniques of evaluation; it is because the discipline will probably always seem unimportant until it (or its neglect) bites you, and quick education about *that particular* branch or application of evaluation will then become very important. No professional who is unsophisticated about personnel, product, proposal and program evaluation in their field *is* a professional; but even when (or if) this sophistication is widespread, application to oneself and one's own programs will not be easy, and the evaluator can help to teach one how to handle the process and its results. When Socrates said, "The unexamined life is not worth living," he was identifying himself as an evaluator; but it is not *accidental* that he is best-known as a teacher. Nor is it accidental that he was killed for combining the two roles. See also **Value-phobia.**

EDUCATIONAL OR OVERALL SIGNIFICANCE To get this, the evaluator must examine the data corresponding to each of the prior checkpoints on the Key Evaluation Checklist: educational significance represents a *total synthesis* of all you know. In particular, the gains attributed to the program or product being evaluated, must be *educationally* significant/valuable and not just be **statistically significant,** something which may only be the result of using a large sample, or due to irrelevant vocabulary gains, poor test construction, peculiar statistical analysis or some other insignificant variable. (The same applies for medically significant, socially significant, etc.)

EFFECTIVENESS Usually refers to *goal-achievement.* Various indexes of effectiveness were developed around mid-century, when evaluation was thought of as simply goal-achievement measurement for social action programs.

EIR See **Environmental Impact Report.**

ENEMIES LIST Worst enemies often make best critics. They have two advantages over friends, in that they are more *motivated* to prove you wrong, and more *experienced*

with a radically different viewpoint. Hence they will often probe deep enough to uncover assumptions one has not noticed, and destroy complacence about the impregnability of one's inferential structures. Obviously we should use them for metaevaluation, and pay them well. But who enjoys working with, thanking, and paying their enemies? The answer is: A good evaluator. This is a key test of the "evaluation attitude" (see **Evaluation Skills**.) How little we really care about the correct assessment of merit and how much we prefer to make life easy for ourselves shows up nowhere more clearly than on this issue. A good example is the distribution of teaching-evaluation forms to students in a college class, normally done near the end of the semester. But where are your enemies then? Long gone; only the self-selected remain. You should distribute the forms to every warm body that crosses the threshold on the first day and any later date; to be turned in to their seat-neighbor when they decide not to come back. It is the ones who left who can tell you the most—by now you know most of what the stalwarts will say. If you value quality, reach out for suggestions to those who think you *lack* it.

ENGINEERING MODEL See **Medical Model.**

ENJOYMENT Although it is an error in educational evaluation to treat enjoyment as primary and learning as not worth direct inspection, there's no justification for not counting enjoyment at all (Kohlberg once commented on the big early childhood program evaluations that it was too bad no one bothered to check whether at least the kids cried less in Headstart centers than at home.) And the situation in certain cases, e.g. aesthetic education, is much nearer to one where enjoyment is a primary goal. A common fallacy is to argue that since it would be a serious mistake to teach K-3 children some cognitive skills at the expense of making them hate school, we should therefore make *sure* they *enjoy* school and *try* to teach them skills. That prioritization of effort reduces the already meager interest in teaching something valuable, and has never been validated for gains in positive attitude towards school. The teacher is in conflict of interest here, since finger-painting takes less preparation than spatial skill-building.

ENTHUSIASM EFFECT See **Hawthorne Effect.**

ENVIRONMENTAL IMPACT REPORT (EIR) Often required by law prior to granting building or business permits or variance. A form of evaluation focusing on the ecosystem effects. Currently based mainly on bio-science and/or traffic analysis, these tend to be thin on the evaluation of opportunity costs, indirect costs, ethics contingency trees, etc.

ERRORS OF MEASUREMENT It is a truism that measurement involves some error; it is more interesting to notice exactly how these errors can get one into trouble in evaluation studies. For example, it is obvious that if we select the low scorers on a test for remedial work, then some of these will be in the group because of errors of measurement (i.e. their performance on the particular test items that were used does not give an accurate picture of their ability). It follows that a remeasurement, using a test of matched difficulty, would immediately place them somewhat higher. Hence on a posttest, which is essentially such a retesting, they will come out looking better, although in fact this is not due to any merit of the intervening treatment, but is simply a statistical artefact due to errors of measurement (specifically, a regression effect). It also follows that matching two groups for their entry level skills, where we plan to use one of them as the control in a *quasi*-experimental study (i.e. one where the two groups are not created by random assignment) will get us into trouble because the errors of measurement on the two groups cannot be assumed to be the same, and hence the regression effect will be different in size. Another nasty effect of errors of measurement is to reduce correlation coefficients; one may intuitively feel that if the errors of measurement are relatively random, they should "average out" when one comes to look at the correlations, but the fact is that the larger the errors of measurement, the smaller the correlations will appear. See **Regression to the Mean.**

ESCROW A neutral individual or secure place where identifying data can be deposited until completion of an evaluation and/or destruction. (Term originated in the law.) See **Filter**, **Anonymity**.

ETHICS (in evaluation) See also **Responsibility Evaluation.** Ethics is the ultimate normative social science, ulti-

mate because it refers to duties (etc.) which transcend all other obligations such as those to prudence, science, and professionalism. It is in one sense a branch of evaluation, in another a discipline which, like history or statistics, contributes a key element to many evaluations. That it is (logically) a social science is of course denied by virtually all social scientists, who have valuephobia about even the suggestion that non-ethical value-judgments have a place in science and hypervaluephobia about importing ethical judgments. But the inexorable consequence of the development of game- and decision-theory, latent function analysis, democratic theory in political science, welfare economics, analytical jurisprudence, behavioral genetics, and the "good reasons" approach to ethical theory, is that all the bricks have been baked for the building, and it's just superstitious to argue that some mysterious force prohibits putting one on top of another. The Constitution and Bill of Rights are essentially ethical propositions, with two properties: first, there are good reasons for adopting them; second, they generate sound laws. The arguments for them (e.g. Mill's "On Liberty") are as good social science as you'll find in a long day's walk through the professional journals, and the inferences to specific laws are well-tested. It follows that all the well-known arguments for law and order are indirectly arguments for the (secular) ethics of the Constitution and for the axiom of equal rights from which they flow, just as the arguments for the existence of atoms are, indirectly, arguments for the existence of electrons. Ethics is just a general social strategy and no more immune to criticism by social science than the death penalty or excise taxes or behavior therapy or police strikes. To act as if some logical barrier prevents science from arguing for or against particular ethical claims such as the immorality of the death penalty, a question of overall *social* strategy, but not from arguing for or against particular strategies within economics or penology is to cut the social sciences off from the most important area in which they can make a social contribution. And it leads to ragged edges on and inconsistencies within the sciences themselves. For an excellent discussion of the "ethics-or-else" dilemma for allocation theory, see E.J. Mishan, ***Cost-Benefit Analysis,*** 1976, Praeger, Chapter 58, "The Social Rationale of Welfare Economics." Interest-

ingly enough, although a large part of that book is about evaluation, (e.g. Chapter 61 is called "Consistency in Project Evaluation,") neither that term nor the author's frequently-used variation "valuation" gets into the index. See **Valuephobia.**

EVALUABILITY Projects and programs—and the plans for them—are beginning to be scrutinized quite carefully for evaluability. This might be thought of as the first commandment of **accountability** or as a refinement of Popper's requirement of *falsifiability*. The underlying principle can be expressed in several ways, e.g. "It is not enough that good works be done, it must be possible to *tell* that (and, more importantly, when) good works have been done." Or "You can't learn by trial and error if there's no clear way to identify the errors." The bare requirement of an evaluation component in a proposal has been around for a while; what's new is a more serious effort to make it feasible and appropriate. That presupposes more expertise in evaluation than most review panels and project monitors have; but that *may* come. Evaluability should be checked and improved at the planning and **preformative** stages. Requiring evaluability of new programs is analogous to requiring *serviceability* in a new car; obvious enough, but who besides fleet owners (and GSA)knew that there was for many years a 2:1 difference in standard service costs as between Ford and GM? Congress may some day learn that low evaluability has a high price.

EVALUAND Whatever is being evaluated; if it is a person, the term "evaluee" is more appropriate.

EVALUATION The process of determining the merit or worth or value of something; or the product of that process. The special features of evaluation, as a particular kind of investigation (distinguished e.g. from traditional empirical research in the social sciences), include a characteristic concern with cost, comparisons, needs, ethics, and its *own* political, ethical, presentational, and cost dimensions; and with the supporting and making of sound value judgments, rather than hypothesis-testing. The term is sometimes used more narrowly (as is "science") to mean only systematic and objective evaluation, or only the work of people labeled "evaluators." While evaluation in the broad sense is ines-

capable for rational behavior or thought, professional evaluation is frequently worthless and expensive. Evaluation—properly done—can be said to be "a science" in a loose sense, as can, for example, teaching; but it is also an art, an inter-personal skill, something that judges and juries and literary critics and real estate assessors and jewelry appraisers do—and thus not "one of the sciences." See also **Formative/Summative, Analytical/Holistic,** etc.

EVALUATION EDUCATION Consumer education is still rather weak on training in evaluation, which should be its most important component. And of course there are other contexts than those in which one's role is that of the consumer, where evaluation education would be most valuable, notably the manager role, or the service-provider / professional role. Few teachers, for example, have the faintest idea how to evaluate their own work, although this is surely the minimum requirement of professionalism. The last decades have seen considerable federal and state effort to provide reasonable standards of quality that will protect the consumer in a number of areas; they have not yet really understood that the superimposition of standards is a poor substitute for understanding the justification for them. Evaluation *training* is the training of (mainly professional) evaluators; evaluation education is the training of the citizenry in evaluation techniques, traps, and resource-finding, and is the only satisfactory long-run approach to improving the quality of our lives without extraordinary wastage of resources.

EVALUATION ETHICS AND ETIQUETTE Because evaluation in practice so often involves tricky interpersonal relations it has much to learn from diplomacy, arbitration, **mediation,** negotiating, and management (especially personnel management). Unfortunately, the wisdom of these areas is poorly encapsulated into learning and training materials, which are mainly truistic or anecdotal. The correct approach would appear to be via the refinement of normative principles and the collateral development of extensive calibration examples, rather as in developing skill in applied ethical analysis (casuistry.) An example: you are the only first-timer on a site-visit team to a prestigious institution, and you gradually realize, as the time slips away in socializing and reading or listening to reports from adminis-

trators and administration-selected faculty, that no serious evaluation is going to occur unless *you* do something about it. What should you do? There is a precise (flow-chartable) solution which specifies a sequence of actions and utterances, each contingent upon the particular outcome of the previous act, and which avoids unethical behavior while minimizing distress; mature professionals without evaluation experience never get it right; some very experienced and thoughtful evaluators come very close; a group containing both reaches complete consensus on it after a twenty-minute disussion. Like so much in evaluation, this shows it to meet the standards of common-sense though it is not in our individual repertoires. It should be. Another example: a write-in response on an anonymous personnel evaluation form accuses the evaluee of sexual harassment. As the person in charge of the evaluation, what *exactly* should you do? ("Ignore it" is not only ethically wrong, it is obviously impossible.)

EVALUATION OF EVALUATIONS. See **Meta-evaluation.**

EVALUATION OF EVALUATORS Track record, not publications, is the key, but how do you get it? See **Evaluation Registry.**

EVALUATION PREDICATES The distinctively evaluative relations or ascriptions involved in **grading, ranking, scoring,** and **apportioning.**

EVALUATION REGISTRY A concept half-way to the certification or licensing of evaluators from complete laissez-faire. This would operate by encouraging evaluators and their clients to file a copy of their joint contract or letter of agreement with the evaluation registry at the beginning of an evaluation; to this would be appended any modifications made along the way and finally a brief standard report by each party, made independently, assessing the quality and utility of the evaluation, and the performance of the client. Each would have a chance to add a brief reaction to the other's evaluation, and the net end result (2 pages) would then be available for inspection, for a fee, by potential clients. This arrangement, it is argued, would be of more use to the client than asking an evaluator to suggest former clients as references or simply looking at a list of publica-

tions or reports, but would avoid the key problems with licensing—enforcement standards, and funding. Start-up costs for such a registry, although small, are not available, possibly because we are in a period of evaluation backlash.

EVALUATION RESEARCH Evaluation done in a serious scientific way; the term is popular amongst supporters of the **social science model** of evaluation.

EVALUATION-SPECIFIC METHODOLOGY Much of the methodology used in evaluation studies is derived from other disciplines—the special nature of evaluation is the way in which it synthesizes these into an appropriate overall perspective, and brings them to bear on the various kinds of evaluation tasks. But there are some situations where essential variations on the usual procedures in scientific research become appropriate. Two instances will be mentioned. In survey research, sample size is normally predetermined in the light of statistical considerations and prior evidence about population parameters. In evaluation, although there are occasions when a survey of the classical kind is appropriate, surveys are frequently *investigatory* rather than *descriptive* surveys and then the situation is rather different. Suppose that a respondent, in a phone interview evaluation survey of users of a particular service, comes up with a wholly unexpected comment on the service which suggests—let us say—improper behavior by the service-providers. (It might equally well suggest an unexpected and highly beneficial side-effect.) This respondent is the thirtieth interviewee, from a planned sample of a hundred. On the standard survey pattern, one would continue, using the same interview form, through the rest of the sample. In evaluations, one will quite often want to alter the form so as to include an explicit question on this point. Of course, one can no longer report the results of the survey with a sample of a hundred, with respect to this question (and any others with whom its presence might interact). But one may very well be able to turn up another twenty people that respond under cueing, who would not have produced this as a free response. That result is much more important than salvaging the survey—in most cases. It also points to another feature of the evaluation situation, namely the desirability of time-sequencing the interviews or question-

naire responses. Hence one should try to avoid using a single mass mailing, a common practice in survey research; by using sequential mailing one can examine the responses for possible modifications of the instrument. The second taboo that we may have good reason to break concerns sample size. If we find ourselves getting a very highly standard kind of response to a fairly elaborate questionnaire, we are discovering that the population has less variability than we had expected, and we should alter our estimate of an appropriate sample size in mid-stream. No point in continuing to fish in the same waters if you don't get a bite after an hour. The generalization of this point is to the use of "emergent", "cascading", or "rolling" designs, where the whole design is varied en route as appropriate. (These terms come from the glossary in ***Evaluation Standards.***) Other evaluation-specific methodology includes the use of parallel teams working independently, **calibration** of judges, **convergence** sessions, "blind" judges, synthesis, bias balancing etc. See also **Anonymity, Questionnaires.**

EVALUATION SKILLS There are lists of desirable skills for evaluators (Stufflebeam has one with 234 competencies); as for philosophers, almost any kind of specialized knowledge is advantageous, and the more obvious tool skills alone (see the **Key Evaluation Checklist**) are far more demanding than in any other discipline—statistics, cost-analysis, ethical analysis, management, teaching, therapy, contract law, graphics, synthesis, dissemination (for the report); and of course there are the **evaluation-specific** techniques. Here we mention a couple that are less obvious. First, the evaluative attitude or temperament. Unless you are committed to the search for quality, as the best of those in other professions are committed to the search for justice or the search for truth, you are in the wrong game. You will be too easily tempted by the charms of "joining" (e.g. joining the program staff—see **Going Native**); too unhappy with the outsider's role. The virtue of evaluation *must be* its own real reward, for the slings and arrows are very real. (Incidentally, this value is a learnable and probably even a teachable characteristic for many people; but some people come by it naturally and others will never acquire it.) The second package of relatively unproclaimed skills are "practical logical analysis" skills e.g. identifying hidden agendas

or unnoticed assumptions about the dissemination process, or mismatches between a goal-statement and a needs-statement, or loopholes in an evaluation design; the ability to provide accurate summaries one fiftieth of the length of the original (precis) or to give a totally non-evaluative, non-interpretive description of a program or treatment. The good news is that no-one is good at all the above; that there is room for specialists, and also for team members. Partly because of the formidable nature of the relevant skills list, evaluation is a field where teams *if properly employed* are immensely better than soloists. Not only are two heads better than one, six (carefully chosen and appropriately instructed) are better than five.

EVALUATION STANDARDS A set of principles for the guidance of evaluators and their clients. The major effort is the ***Evaluation Standards*** (ed. D. Stufflebeam, McGraw Hill, 1980), but the Evaluation Research Society has also produced a set. There are some shared weaknesses—for example, neither includes **needs assessments**—but the former is much more explicit about interpretation, giving specific examples of applications etc. In general, these are likely to do good by raising clients' consciousness and general performance, but fears have been expressed by first-rank evaluators that they may rigidify approaches, stifle research, increase costs (cf. "defensive lab tests" in medical practice today), and give a false impression of sophistication. See also **Bias.**

EVALUATION, THEORY OF The theory of evaluation includes a wide range of topics from the logic of evaluative discourse, general accounts of the nature of evaluation and how it can be justified (axiology), through socio-political theories of its role in particular types of environment, to so-called "models" which are often simply conceptualizations of or procedural recommendations for evaluation. Little work is funded on this; a notable exception is NIE's Research on Evaluation project at NWL, a series of studies on radically different "metaphors" for evaluation.

EVALUATION TRAINING There is essentially no serious support for this at the moment, despite the large demand (and larger need) for trained evaluators, perhaps a sign of evaluation backlash. The best places are probably

CIRCE and the Evaluation Center at Western Michigan with post-doc work at NorthWest Labs. Short courses are more widely available and advertised in ***Evaluation News.*** See also **Training of Evaluators.**

EVALUEE A person being evaluated; the more general term, which covers products and programs, etc., is "evaluand."

EXECUTIVE SUMMARY Abstract of results from an evaluation, in non-technical language.

EXIT INTERVIEWS Interviews with subjects as they leave the e.g. training program, clinic, etc., to obtain factual and judgmental data. A very good time for these, with respect to course or teaching evaluation in the school or college setting, is at the time of graduation, when (a) the student will have some perspective on most of the educational experience; (b) fear of retribution is low; (c) response rate can be nearly 100% with careful planning; (d) judgments of effects are relatively uncomplicated. Later than this—alumni surveys—conditions can and do deteriorate, though there is a partial offset because job-relevance can be judged more accurately.

EXPERIMENT See **True Experiment.**

EXPERIMENTAL GROUP The group (or single person, etc.) that is receiving the treatment being studied.

EXPLANATION By contrast with evaluation, which identifies the value of something, explanation involves answering a Why or How question about it, or other type of request for understanding. Often explanation involves finding the *cause* of a phenomenon, rather than its *effects* (which is a major part of evaluation). When it is possible, without jeopardizing the main goals of what may be holistic summative evaluation, a good evaluation design tries to uncover micro-explanations (e.g. by identifying those components of the curriculum package which are producing the major part of the effects, and which are having little effect). The first priority, however, is to resolve the evaluation issues (Is the package the best available? etc.), but too often the research orientation and training of evaluators leads them to do a poor job on evaluation because they got in-

terested in explanation (*LE*). The realization that the logical *nature* and investigatory demands of evaluation are quite different from those of explanation is as important as the corresponding realization with respect to prediction and explanation, which the neo-positivist philosophers of science still think are logically the same under the (temporal) skin.

EX POST FACTO DESIGN One where we identify a control group "after the fact," i.e., after the treatment has occurred. A very much weaker design than the **true experiment** since there must have been *something* different about the subjects that got the treatment without being assigned to it, in order to explain why they got it, and that something means they're not the same as the control group, in some unknown respect that may be related to the treatment.

EXTERNAL (evaluator or evaluation) An external evaluator is someone who is *at least* not on the project or program regular staff, or someone—in the case of personnel evaluation—other than the individual being evaluated, or their staff. It is *better* if they are not even paid by the project or by any entity with a prior preference for the success or failure of the project. Where or to whom the external evaluator reports is what determines whether the evaluation is formative or summative, *either* of which may be done by external or by internal evaluators (contrary to the common view that external is for summative, internal for formative), and both of which should be done by both.

EXTERNAL VALIDITY By contrast with **internal validity,** this refers to the generalizability of the experimental/evaluation findings. Here the traps to avoid include failure to identify key environmental variables that happen to be constant throughout the experiment, decreased sensitivity of participants to treatment at posttest due to pretest, reactive effects of experimental arrangement, or biased selection of participants that might affect the generalizability of the treatment's effect to non-participants—thus jeopardizing the external validity. (Ref. ***Experimental and Quasi-Experimental Designs for Research,*** D.T. Campbell and J.C. Stanley, Rand McNally & Co., Chicago, 1972.)

EXTRAPOLATE Infer conclusions about ranges of the variables beyond those measured. Cf. **Interpolate.**

FACE VALIDITY The apparent validity, typically of test items or of tests; there can be skilled and unskilled judgments of face validity, and highly skilled judgments which come pretty close to **content validity**, which does require systematic substantiation.

FADING Technique used in programmed texts, where a first answer is given completely, the next one in part with gaps, then with just a single cue, then called for without help. A key technique in training and calibrating evaluators.

FAULT TREE ANALYSIS (CAUSE TREE ANALYSIS) These terms emerged about 1965, originally in the literature of management science and sociology. They are sometimes used in a highly technical sense, but are useful in a straightforward sense. Basically, the model to which they refer is the trouble-shooting chart, often to be found in the pages of e.g. a Volkswagen manual. The branches in the tree identify possible causes of the fault (hence the terms "cause" and "fault" in the phrase), and this method of representation—with various refinements—is used as a device for management consultants, for management training, etc. Its main use in evaluation is as a basis for **needs assessment.**

FIELD INITIATED This refers to proposals or projects for the funding of grants or contracts that originate from workers in the field of study, rather than from a program announcement of the availability of funds by an agency for work in a certain area (which is known as "solicited" research or development.)

FIELD TRIAL (OR FIELD TEST) A dry run of a test of a product/program, etc. Absolutely mandatory in any serious evaluation or development activity. It is essential that at least one true field trial should be done in circumstances and with a population that matches the targeted situation and population. *Earlier* ("hothouse") trials may not meet this standard, for convenience reasons, but the last one must. Unless run by external evaluators (very rare), there is a major risk of bias in the sample or conditions or content or interpretations used by the developer in the final field trials.

FILTER Someone who—or a computer which—removes identifying information from evaluative input, to

preserve the **anonymity** of the respondent.

FISCAL EVALUATION The highly developed subfield that involves looking at the worth or probable worth of e.g. investments, programs, companies. See **ROI, Payback, Time Discounting, Profit,** etc.

FISHING Colloquialism for exploratory (phase of) research; *or* for true nature of large slices of serious (e.g. program) evaluation; or for visits to Washington in search of funding support.

FLOW CHART A graphic representation of the sequence of decisions, including contingent decisions, that is set up to guide the management of projects (or the design of computer programs), including evaluation projects. Usually looks like a sideways organization diagram, being a series of boxes and triangles ("activity blocks," etc.) connected by lines and symbols that indicate simultaneous or sequential activities/decision points, etc. A **PERT** chart is a special case.

FORMATIVE EVALUATION Formative evaluation is conducted *during* the development or improvement of a program or product (or person, etc.). It is an evaluation which is conducted *for* the in-house staff of the program and normally remains in-house; but it may be *done by* an internal *or* an external evaluator or (preferably) a combination. The distinction between formative and summative has been well summed up in a sentence of Bob Stake's "When the cook tastes the soup, that's formative; when the guests taste the soup, that's summative."

FOUND DATA Data that already exists, prior to the evaluation—contrast is with experimental data or test and measurement data.

FUGITIVE DOCUMENT One which is not published through the public channels as a book or journal article. Evaluation reports have often been of this kind. ERIC (Educational Resources Information Center) has picked up some of these, but since its criteria for selection are so variable and its selection so limited, time spent in searching it is all too often not cost-effective.

FUNDING (of evaluations) Done in many ways, but a

common pattern is described here. The evaluation proposal may be "field-initiated," i.e. unsolicited, or sent in response to (a) a program announcement, (b) an RFP (Request for Proposal), (c) a direct request. Typically (a) results in a grant, (b) in a contract; the former identifies a general charge or mission (e.g. "to develop improved tests for early childhood affective dimensions") and the latter specifies more or less exactly what is to be done, e.g. how many cycles of field tests (and who is to be sample, how large a sampled is to be used, etc.), in a "Scope of Work." The legal difference is that the latter is enforceable for lack of performance, the former is (practically) not. But it scarcely makes sense to use contracts for research (since you usually can't foresee which way it will go), and it is rarely justifiable to use them for the very specific program evaluations required by law. Approach c, "sole-sourcing", eliminates competitive bidding and can usually only be justified when only one contractor has much the best combination of relevant expertise or equipment or staff resources; but it is much faster, and it does avoid the common absurdity of 40 bidders, each spending 12K ($12,000) to write a proposal worth 300K to the winner. The wastage there (180K) comes out of overhead costs which are eventually paid by the taxpayer, or by bidders going broke because of foolish requirements. A good compromise is the **two-tier system,** all bidders submitting a two (or five or ten) page preliminary proposal, the best few then getting a small grant to develop a full proposal. Contracts may or may not have to be awarded to the lowest "qualified" bidder; qualification may involve financial resources, stability, prior performance, etc., as well as technical and management expertise. On big contracts there is usually a "bidders' conference" shortly after publication of the RFP (it's often required that federal agencies publish the RFP in the ***Business Commerce Daily*** and/or the ***Federal Register***). Such a conference officially serves to clarify the RFP; it may in fact be a cross between a con job and a poker game. If you ask clever questions, others may (a) be scared off, (b) steal your approach, etc. The agency may be sniffing around for a "friendly" evaluator and the evaluators may be trying to look friendly but not so friendly as to reduce credibility, etc. Eventually, perhaps after a second bidders' conference, the most promising bidders will be asked for

their Best and Final bid and on this basis the agency selects one, probably using an (anonymous, possibly) external review panel to lend credibility to the selection. After the first conference between the winner and the project officer (the agency's representative used to be called the monitor) it often turns out that the agency wants or can be persuaded to want something done that isn't clearly in the contract; the price will then be renegotiated. Or if the price was too low (the RFP will often specify it in terms of "Level of Effort" as N "person-years" of work; this may mean N × 30K or N × 50K in dollar terms, depending on whether overhead is an add-on) to get the job done, the contractor may just go ahead till they run out of money and then ask for more, on the grounds the agency will have sunk so much in and be so irreversibly committed (time-wise) that they have to come through to "save their investment." The contractor of course loses credibility on later bids but that's better than bankruptcy; and the track-records are so badly kept that no one may hold it against them (if indeed they should). In the bad old days, low bids were a facade and renegotiation on trumped-up grounds would often lead to a cost well above that of another and better bidder. Since evaluations are tricky to do in many ways, bidders have to allow a pad in their budget for contingencies—or just cross their fingers, which quickly leads to bankruptcy. Hence another option is to RFP for the best design and per diem and then let the contract for as long as it takes to do it. The form of abuse associated with this cost-plus approach is that the contractor is motivated to string it out. So no overall clear saving is attached to either approach; but the latter is still used where the agency wants to be able to change targets as preliminary results come in, a sensible point, and where it has good monitoring staff to prevent excessive over-runs (from *estimates* which of course are not binding). A major weakness in all of these approaches is that innovative proposals will often fail because the agency has appointed a review panel of people committed to the traditional approaches who naturally tend to fund "one of their own." Another major weakness is the complexity of all this, which means that big organizations who can afford to open branches in D.C., pay professional proposal-writers and "liaison staff" (i.e., lobbyists), have a tremendous edge (but often do poor work, since most of the best people do no

work for them). A third key weakness is that the system described favors the product of timely paper rather than the solution of problems, since that's all the monitoring and managing process can identify. Billions of dollars, millions of jobs, thousands of lives are wasted because we have no reward system for really good work, that produces really important solutions. The reward is for the *proposal*, not the product; and it is the contract. Once obtained, only unreliability in delivery or gross negligence jeopardizes future awards. You can see the value system this arrangement produces from the way the vice-presidents all move on to work on the next "presentation" as soon as negotiation is complete. It would only cost pennies to reverse this via (partial) contingency awards and expert panels to review work done instead of proposals.

FUTURISM Since many evaluands are designed to serve future populations and not (just) present ones, much evaluation requires estimating future needs and performance. The simpler aspect of this task involves extrapolation of demographic data; even this is poorly done e.g. the crunch on higher education enrolments was only foreseen by one analyst (Cartter) although the inference was simple enough. The harder task is predicting e.g. vocational patterns twenty years ahead. Here one must fall back on possibility-covering techniques, rather than probability-selection e.g. by teaching flexibility of attitude or generalizable skills.

GOAL The technical sense of this term restricts its use to rather general descriptions of intended outcome; more specific descriptions are referred to as **objectives.**

GOAL-ACHIEVEMENT MODEL (of evaluation) The idea that the merit of the program (or person) is to be equated with success in achieving a stated goal. This is the most naive version of **goal-based evaluation.**

GOAL-BASED EVALUATION (GBE) This type of evaluation is based and focused on knowledge of the goals and objectives of the program, person or product. A goal-based evaluation often does not question the merit of goals; often does not look at *cost*-effectiveness; often fails to search

for or locate the appropriate **critical competitors**; often does not search for side effects; in short, often does not include a number of important and necessary components of an evaluation. Even if it does include these components, they are *referenced to* the program's (or to personal) goals and hence run into serious problems such as identifying these goals, handling inconsistencies in them and changes in them over time, dealing with shortfall and overrun results and avoiding the perceptual bias of knowing about them. GBE is *manager-oriented* evaluation, close to **monitoring** and far from *consumer-oriented* evaluation. (See **GFE**).

GOAL-FREE EVALUATION (GFE) In this type of evaluation, the evaluator(s) is not told the purpose of the program but enters into the evaluation with the purpose of finding out what the program actually is *doing* without detailed cueing as to what it is *trying* to do. If the program *is* doing what its stated goals and objectives say, then these achievements should show up (in observation of process and interviews with consumers (not staff)); if not, it is argued, they are irrelevant. Merit is determined by relating program *achievements* to the *needs* of the *impacted* population, rather than to the program (i.e., agency or citizenry or congressional or manager's) goals. It could thus be called "needs-based evaluation" or "consumer-oriented evaluation" by contrast with goal-based or manager-oriented evaluation. It does *not* substitute the evaluator's goals for the program's goals, nor the goals of the consumer; the evaluation must justify (via the needs assessment) all assignments of merit. GFE is generally disliked by both managers/administrators and evaluators, for fairly obvious reasons. It is said to be less intrusive than GBE, more adaptable to mid-stream goal shifts, better at finding side effects and less prone to social, perceptual and cognitive bias. It is risky, because the client may get a nasty shock when the report comes in (no prior hand-holding) and refuse to pay because embarrassed at the prospect of having to pass the evaluation along to funding agency. (But if the findings are invalid, the client should simply document this and ask for modifications.) GFE is reversible, a key advantage over GBE; hence an evaluation design should (sometimes) *begin* GFE, write a preliminary report, then go to GBE to see if serious errors of omission occurred. (Running a *parallel* GFE

effort along with a GBE reduces the time-span.) The shock reaction to GFE in the area of program evaluation (it is the standard procedure used by all consumers evaluating products) suggests that the grip of management bias on program evaluation was very strong, and possibly that managers felt they had achieved considerable control over the outcomes of GBEs.GFE is analogous to double-blind design in medical research; even if the evaluator would like to give a favorable report (e.g. because of being paid by the program, or because hoping for future work from them) it is not (generally) easy to tell *how* to "cheat" under GFE conditions. The risk of failure by an evaluator is of course greater in GFEs, which is desirable since it increases effort, identifies incompetence, and improves the **balance of power.**

GOING NATIVE The fate of evaluators that get co-opted by the programs they are evaluating. (Term originated with the Experimental Schools Progam evaluation in mid-60's.) The co-option was often entirely by choice and well illustrates the pressures on, temptations for, and hence the temperamental requirements for being a good evaluator. It can be a very lonely role and if you start thinking about it in the wrong way you start seeing yourself as a negative force—and who wouldn't rather be a co-author than a (mere) critic? One answer; someone who cares more about quality than kudos. See **Evaluation Skills.**

GRADE-EQUIVALENT SCORE A well-meant attempt to generate a meaningful index from the results of standardized testing. If a child has a 7.4 grade-equivalent score, that means s/he is scoring at the average level (estimated to be) achieved by students four months into the 7th grade. Use of the concept has often led to an unjustified worship of average scores as a reasonable standard for individuals, and to overlooking the **raw scores** which may tell a very different story. Suppose a beginning eighth grader is scoring at the 7.4 level; parents may be quite upset unless someone points out that on this particular test the 8.0 level is the same as the 7.4 level (because of summer backsliding). In reading, a deficit of two whole grade equivalents is quite often made up in a few months in junior high school if a teacher succeeds in motivating the student for the first time. Again, a student may be a whole grade-equivalent down and be

ahead of most of the class—if the average score is calculated as the *mean* not the median. Again, a student in the *fifth* grade scoring 7.2 might flunk the seventh grade reading test completely; 7.2 just means that s/he scores where a seventh grader would score on the *fifth* grade test. A year's deficit from the 5th grade norm isn't comparable to a year's deficit from the 4th grade norm. And so on—i.e., use with caution.

GRADING ("Rating" is sometimes used as a synonym.) Allocating individuals to an ordered (usually small) set of labeled categories, the order corresponding to merit, e.g. A–F for "letter grading." Those within a category are regarded as tied if the letter grade only is used; but if a numerical grade (**"scoring"**) is also used, they may be ranked within grades. The use of plus and minus grades simply amounts to using more categories. Grading provides a *partial* **ranking,** but ranking cannot provide grading without a further assumption, e.g. that the best student is good enough for an A, or that "grading on the curve" is justified. That is, the grade labels normally have some independent meaning and cannot be treated as simply a sequenced set that can be distributed by making arbitrary cuts in a ranked sequence of individuals. In short, the grades are normally criterion-referenced and ranking is normally facilitated by norm-referenced testing: that tension frequently results in confusion. For example, grading of students does not imply the necessity for "beating" other students, does not need to engender "distractive competitiveness" as is often thought. Only *publicized grading on a curve* does that. Pass/Not Pass is a simple form of grading, not a no-grading system. Grades should be treated as quality estimates by an expert and thus constitute essential feedback to the learner or consumer; corrupting that feedback because the external society misuses the grades is abrogation of duty to the learner or consumer. See **Responsibility Evaluation.**

GRANT See **Funding.**

HALO EFFECT The tendency of someone's reaction to part of a stimulus (e.g. part of a test, part of a student's answers to a test, part of someone's personality) to spill over

into their reaction to other, especially adjacent, parts of the same stimulus. For example, judges of exams involving several essay answers will tend to grade the second answer by a particular student higher if they graded the first one high than they would if this had been the first answer they had read by this student (the error is often as much as a full grade). Halo effect is avoided by having judges assess all the first components before they look at any of the second components, and by concealing from them their grade on the first component when they come to evaluate the second one. The halo effect gets its name from the tendency to suppose that someone who is saintly in one kind of situation must be saintly (and perhaps also clever) in all kinds of situations. But the halo effect also refers to the illicit transfer of a *negative* assessment. The Hartshorne & May work (***Studies in Deceit,*** Columbia, 1928) suggests there is no good basis for this transfer.

HARD vs. SOFT (approaches to evaluation) Colloquial way to refer to the differences between the quantitative /testing / measurement / experimental-design approach to evaluation and the descriptive / observational / narrative/ ethnographic / participant-observer kind of approach.

HAWTHORNE EFFECT The tendency of a group or person being investigated, or experimented on, or evaluated to react positively or negatively to the fact that they are being investigated/evaluated, and hence to perform better (or worse) than they would in the absence of the investigation, thereby making it difficult to identify any effects due to the treatment itself. Not the same as the enthusiasm effect, i.e., the effect on the consumer of an enthusiastic service-provider that results simply from the enthusiasm of provider or recipient. The placebo effect is the medical analog of the enthusiasm effect.

HEADROOM See **Ceiling Effect.**

HIERARCHICAL SYSTEM See **Two-Tier.**

HOLISTIC SCORING/GRADING/EVALUATING The allocation of a single score/grade/evaluation to the overall performance of an evaluand; by contrast with **analytical scoring/grading/evaluating.** The holistic/analytical distinction corresponds to the macro/micro distinction in econom-

ics and the molar/molecular distinction in psychology.

HYPERCOGNITIVE or TRANSCOGNITIVE The domain beyond the **supercognitive,** which is the stratosphere of the cognitive; includes meditation and concentration skills; originality; the intellectual dimension of empathic insight (as evidenced in role-playing, acting, etc.); eidetic imaging; near-perfect objectivity, rationality, reasonableness, or "judgment" in the common parlance; moral sensitivity; ESP skills, etc. Some of this is incorrectly included in "affective education."

HYPOTHESIS TESTING The standard model of scientific research in the classical approach to the social sciences, in which a hypothesis is formulated prior to the design of the experiment, the design is arranged so as to test its truth, and the results come out in terms of a probability estimate that only chance was at work. If the probability is extremely low that only chance was at work, the design should make it inductively highly likely that the hypothesis being tested was correct. What is to count as the high degree of improbability that only chance was at work is usually taken to be either the .05 "level of significance" or the .01 "level of significance." When dealing with phenomena whose existence is in doubt, a more appropriate level is .001; where the occurrence of the phenomenon in this particular situation is all that is at stake, the conventional levels are more appropriate. The significance level is thus used as a crude index of the merit of a hypothesis.

An important distinction in hypothesis testing that does carry over to the evaluation context in a useful way is the distinction between Type 1 and Type 2 errors. A Type 1 error is involved when we conclude that the null hypothesis is false although it isn't; a Type 2 error is involved when we conclude that the null hypothesis is true when in fact it's false. Using a .05 significance level means that in about 5% of cases studied, we will make a Type 1 error. As we tighten up on our level of significance, we reduce the chance of Type 1 error, but correspondingly increase the chance of a Type 2 error (and *vice versa*). It is a key part of evaluation to look carefully at the relative costs of Type 1 and Type 2 errors. (In evaluation, of course, the conclusion is about merit rather than truth.) A metaevaluation should carefully

spell out the costs and benefits of the two kinds of error, and scrutinize the evaluation for its failure or success in taking account of these in the analysis, synthesis, and recording phases. For example, in quality control procedures in drug manufacture (a type of evaluation), it may be fatal to a prospective user to identify a drug sample as satisfactory when in fact it is not; on the other hand, identifying it as unsatisfactory when it is really satisfactory will only cost the manufacturer whatever that sample costs the manufacturer to make. Hence it is obviously in the interest of the public and the manufacturer (given the possibility of damage suits) to set up a system which minimizes the chance of false acceptances, even at the expense of a rather high level of false rejections. Because of the totally non-mnemonic characteristics of the terms "Type 1" and "Type 2," it's always better to use terms like "incorrect acceptance" and "incorrect rejection" of *evaluands*, rather than of the *null hypothesis*, the latter concept being likely to prove unenlightening to most audiences.

ILLUMINATIVE EVALUATION (Rippey) A type of pure process evaluation, very heavy on multi-perspective description and interpersonal relations, very light on justified tough standards, very easy on **valuephobes.**

IMPACT EVALUATION An evaluation focussed on outcomes or pay-off rather than process delivery or implementation evaluation.

IMPACTED POPULATION The population that is crucial in evaluation, by contrast with the target population and even the **true consumers.**

IMPLEMENTATION EVALUATION Recent reactions to the generally unexciting results of impact evaluations on social action programs have included a shift to mere monitoring of program delivery i.e. implementation evaluation. You can easily implement; it's harder to improve.

IMPLEMENTATION OF EVALUATIONS The frequent complaint that evaluations have little effect, i.e. are not implemented, refers to four quite different situations. (a) Many evaluations are simply incompetent and it's most

desirable they *not* be implemented; (b) Some evaluations make—and should make—no immediate recommendations (e.g. accountability evaluations); nevertheless they have a powerful preventive effect and some cumulative long-run effect, but neither is readily measurable; (c) Many evaluations are commissioned in such a way that even when done as well as possible they will not be of any use because they were set up so as to be irrelevant to the real issues that affect the decision-maker, or one so underfunded that no sound answer can be obtained—again, it is just as well these not be implemented; (d) Some excellent evaluations are ignored because the decision-maker doesn't like (e.g. is threatened by) the results or won't take on the risks or trouble of implementation. The lack of implementation phenomenon thus has little or large implications for the field of evaluation, depending entirely on the distribution of the causes across these four categories. It is hardly something to be unduly concerned about professionally as long as evaluation still has a long way to go in doing its own job well; doctors shouldn't worry that their patients ignore their advice if it's bad. But as a *citizen* one can scarcely avoid worry about the colossal wastage resulting from the fourth kind of situation; here's a fairly typical quote from the 8/1/80 GAO reports on their (usually very good) evaluations: "The Congress has an excellent opportunity to save billions of dollars by limiting the number of noncombat aircraft to those that can be adequately justified . . . Dept. of Defense justifications [were] . . . based on unrealistic data and without adequate consideration of more economical alternatives." GAO has been issuing reports on this topic since 1976 without noticeable effect so far.

IMPLEMENTATION OF TREATMENT The degree to which a treatment has been instantiated in a particular situation, typically a field trial of the treatment or an experimental investigation of it. The notion of an "index of implementation," consisting of a set of scales describing the key features of the treatment, and allowing one to measure the extent to which it is manifested in each dimension, is a useful one for checking on implementation, an absolutely fundamental check if we are to find out whether the treatment has merit. This is part of the "purely descriptive" effort in evaluation, and is handled under the description

checkpoint and the process checkpoint of the **Key Evaluation Checklist.** One characteristic situation occurs when the description checkpoint provides a correct account of the treatment that is supposed to be implemented, and the process checkpoint provides a correct description of what is actually occurring; the match between the two is a measure of the implementation, and hence of the extent to which we can generalize from the results of the test to an evaluation of the evaluand which we are *supposed* to be evaluating.

INCESTUOUS RELATIONS (in evaluation) Refers to (a) extreme **conflict of interest** (where the evaluator is "in bed with" the program being evaluated), as is typical of ordinary program monitoring by agencies and foundations where the monitor is usually the godfather (sic) of the program, sometimes its inventor and nearly always its advocate at the agency, and a co-author of its modifications as well as—supposedly—its evaluator; (b) incestuous validation of test items occurs when they are selected/rejected on the basis of the correlation of performance on that item with overall score on the test. Many widely-used tests have lowered their construct validity by dumping face-valid items because of this. The correct procedure is to check for other errors (e.g. irrelevance, ambiguity) perhaps by external judge review or rewriting the item(s), hoping the correlation won't hold up—because then you have tapped into an independent dimension of criterion performance.

INCREMENTAL NEED An unmet **need.**

INDEPENDENCE Independence is only a relative notion; but by increasing it, we can decrease certain types of bias. Thus, the external evaluator is somewhat more independent than the internal, the consulting medical specialist can provide a more "independent opinion" than the family physician, and so on. But of course both may share certain biases, and there is always the particular bias that the external or "second opinion" is typically hired by the internal one and is thus dependent upon the latter for this or later fees, a not inconsiderable source of bias. The more subtle social connections between members of the same profession, e.g. evaluators, are an ample basis for suspicion about the true independence of the second or **meta-evaluator**'s opinion. The best approach is typically to use more than one

"second opinion" and to sample as widely as possible in selecting these other evaluators, hoping from an inspection of their (independently written) reports to obtain a sense of the variation within the field, from which one can extrapolate to an estimate of probable errors.

INDEPENDENT VARIABLE See **Dependent Variable.**

INDICATOR A factor, variable, or observation that is empirically *or* definitionally connected with the criterion; a correlate. For example, the judgment by students that a course has been valuable to them for pre-professional training is a (weak) indicator of that value. Criteria, by contrast, *are*, or are *definitionally* connected with, the "criterion" (real pay-off) variable. Indicators thus include but are not limited to criteria. *Constructed* indicators are variables *designed* to reflect e.g. the health of the economy (a social indicator) or the effectiveness of a program. They, like course grades, are examples of the frequent need for concise evaluations even at the cost of some accuracy and reliability.

INFERENTIAL STATISTICS That part concerned with making inferences from characteristics of samples to characteristics of the population from which the sample comes, which of course can only be done with a certain degree of probability (cf. **Descriptive Statistics**). Significance tests and confidence intervals are devices for indicating the degree of risk involved in inference (or "estimate")—but they only cover some dimensions of the risk. For example, they cannot measure the risk due to the presence of unusual and possibly relevant circumstances such as freakish weather, an incipient gas shortage, ESP, etc. Judgment thus enters into the final determination of the probability of the inferred condition.

INITIATION-JUSTIFICATION BIAS See **Consonance Dissonance.**

INFORMAL LOGIC Several evaluation theorists consider evaluation to be in some respects or ways a kind of persuasion or argumentation (notably Ernest House, in ***Evaluating with Validity***, Sage, 1980). In terms of this view, it is relevant that there are new movements in logic, law and science which give more play to what have previously been dismissed as "merely psychological" factors e.g. feelings,

understanding, plausibility, credibility. The "informal logic movement" parallels that of the **New Rhetoric** and *naturalistic* methodology in the social sciences. Ref. ***Informal Logic*** ed. Johnson and Blair, Edgepress, 1980.

INFORMED CONSENT The state which in conscious adults represents a good start toward discharging one's ethical obligations towards human subjects. The tough cases involve semi-rational semi-conscious semi-adults.

INSTITUTIONAL EVALUATION A complex evaluation, typically involving the evaluation of a set of programs provided by an institution plus an evaluation of the overall management, publicity, personnel policies and so on of the institution. The **accreditation** of schools and colleges is essentially institutional evaluation, though a very poor example of it. One of the key problems with institutional evaluation is whether to evaluate in terms of the mission of the institution or on some absolute basis. It seems obviously unfair to evaluate an institution against goals that it isn't trying to achieve; on the other hand, the mission statements are usually mostly rhetoric and virtually unusable for generating criteria of merit, and they are at least potentially subject to criticism e.g. because of inappropriateness to need of clientele, internal inconsistencies, impracticality with respect to the available resources, ethical impropriety, etc. So one must in fact evalauate the goals *and* the performance relative to these goals or do **goal-free evaluation.** Institutional evaluation always involves more than the sum of the component evaluations; for example, a major defect in most universities is departmental dominance, with the attendant costs in rigidifying career tracks, virtually eliminating the role-model of the generalist, blocking new disciplines or programs—and preserving outdated ones—(since in steady-state they have to come out of the department's budget) etc. Most evaluations of schools and colleges fail to consider these system features, which may be more important than any components.

INTERACTIVE (evaluation) One in which the evaluees have the opportunity to react to the content of a first draft of an evaluative report, which is reworked in the light of any valid criticisms or additions. A desirable approach whenever feasible, as long as the evaluator has the courage

to make the appropriate criticisms and stick to them unless they are repudiated. Very few have, as one can see by looking at site-visit or personnel reports that are not confidential, by comparison with those that are, e.g. verbal supplements by the site visitors.

INTERACTION Two factors or variables interact if the effect of one, on the phenomenon being studied, depends on the magnitude of the other. For example, math education interacts with age, being more or less effective on children depending on their age; and it interacts with math achievement. There are plenty of interactions between variables governing human feelings, thought and behavior but they are extremely difficult to pin down with any precision. The classic example is the search for aptitude-treatment or trait-treatment interactions in education; everyone knows from their own experience that they learn more from certain teaching styles than from others, and that other people do not respond favorably to the same styles. Hence there's an interaction between the teaching style (treatment) and the learning style (aptitude) with regard to learning. But, despite all our technical armamentarium of tests and measuring instruments, we have virtually no solid results as to the size or even the circumstances under which these ATI's occur. (Ref: ***The Aptitude-Achievement Distinction,*** ed. D. R. Green, McGraw Hill, 1974.)

INTERNAL Internal evaluators (or evaluations) are (done by) project staff, even if they are special evaluation staff, i.e., even if they are external to the production/writing/teaching/service part of the project. Usually, internal evaluation is part of the formative evaluation effort, but long term projects have often had special summative evaluators on their staff, despite the low credibility (and probably low validity) that results. Internal/external is really a difference of degree rather than kind; see **Independence.**

INTERNAL VALIDITY The kind of validity of an evaluation or experimental design that answers the question: "Does the design prove what it's supposed to prove about the treatment *on the subjects actually studied?"* (cf. **External Validity**). In particular, does it prove that the treatment produced the effect in the experimental subjects? Relates to the *CAUSATION* checkpoint in the **Key Evaluation Check-**

list. Common threats to internal validity include poor instruments, participant maturation, spontaneous change, or assignment bias. (Ref. ***Experimental and Quasi-Experimental Designs for Research,*** D.T. Campbell and J.C. Stanley, Rand McNally & Co., Chicago, 1972.)

INTEROCULAR DIFFERENCES Fred Mosteller, the great practical statistician, is fond of saying that he's not interested in statistically significant differences, but only in interocular ones—those that hit you between the eyes. Or that's what he's *said* to be fond of saying.

INTERPOLATE Infer to conclusions about values of the variables within the range sampled. Cf. **Extrapolate.**

INTERRUPTED TIME SERIES A type of **quasi-experimental design** in which the treatment is applied and then withheld in a certain pattern, *to the same subjects.* The somewhat ambiguous term "self-controlled" used to be used for such cases, since the control group is the same as the experimental group. The simplest version is of course the "aspirin for a headache" design; if the headache goes away, we credit the aspirin. On the other hand, "psychotherapy for a neurosis" provides a weak inference because the length of the treatment is so great that the chance of the neurosis ending during that interval for other reasons than the psychotherapy is very significant. (Hence short-term psychotherapy is a better bet, *ceteris paribus.*) The next fancier self-controlled design is the so-called "ABBA" design, where A is the treatment, B the absence of it—or another treatment. Measurements are made at the beginning of each labeled period and at the end. Here we may be able to control for the spontaneous remission possibility and sundry interaction effects. This is quite a good design for experiments on supportive or incremental treatments, e.g. we teach 50 words of vocabulary by method A, then 50 more by method B—and to eliminate the possibility that B only works when it follows A, we now reverse the order, and apply it first, and then A. The classic fallacy in this area is probably that of the Governor of Connecticut who introduced automatic license suspension for the first speeding violation and got a very large reduction in the highway fatality rate immediately, about which he crowed a good deal. But a look at the variability of the fatality rate in

previous years would have made a statistician nervous, and sure enough, it soon swung up again in its fairly random way. (Ref. ***Interrupted Time Series Designs,*** Glass, *et al.*, University of Colorado.)

JOB ANALYSIS A breakdown of a job into functional components, often necessary in order to provide remedial recommendations and a framework for micro-evaluation or needs assessment. Job analysis is a highly skilled task, which, like programming, is usually done badly by those hired to do it because of the failure of the pay scale to reflect the pay-offs from doing it well.

JOHN HENRY EFFECT (Gary Saretsky's term) The correlative effect to, or in an extended sense a special case of, the Hawthorne effect, i.e., the tendency of the *control* group to behave differently just because of the realization that they *are* the control group. For example, a control group of teachers using the traditional math program that is being run against an experimental program may—upon realizing that the honor of defending tradition lies upon *them*—perform much better during the period of the investigation than they would have otherwise, thus yielding an artificial result. One cannot of course assume that the Hawthorne effect (on the experimental group) cancels out the John Henry effect.

JUDGMENT It is not accidental—though it was erroneous—that the term "value judgment" came to be thought of as the paradigm of evaluative claims; judgment is a very common part of evaluation, as it is of all serious scientific inference. The function of the discipline of evaluation can be seen as largely a matter of reducing the element of judgment in evaluation, *or* reducing the element of arbitrariness in the necessary judgments e.g. by reducing the sources of bias in the judges e.g. by using double-blind designs, teams, *parallel* teams, **convergence** sessions, **calibration** training etc. The most important fact about judgment is not that it isn't as objective as measurement but that one can distinguish good judgment from bad judgment (and train good judges.)

JUDICIAL OR JURISPRUDENTIAL MODEL (of evaluation) Wolf's preferred term and a term sometimes used

for his version or, rather, extension of **advocate-adversary** evaluation. He emphasizes that the law as a metaphor for evaluation involves much more than an adversarial debate, e.g. the fact-finding phase, cross-examination, evidentiary and procedural rules, etc. It involves a kind of inquiry process that is markedly different from the social scientific one, one that in several ways is tailored to needs more like those of evaluation (the action-related decision, the obligatory simplifications because of time, budget and audience limitations, the dependence on a particular judge and jury, the fate of individuals at stake, etc.). Wolf sees the **educational role** of the judicial process (teaching the jury the rules of just inquiry) as a key feature of the judicial model and it is certainly a strong analogy with evaluation.

JURY TRIAL Used in TA and evaluation. See preceding entry.

KEY EVALUATION CHECKLIST (KEC) What follows is not intended to be a full explanation of the key evaluation checklist and its application, something which would be more appropriate for a monograph on the methodology of evaluation. It simply serves to identify the many dimensions that must be explored prior to the final synthesis in an evaluation. The most important of these are given italicized headings in the checklist, but all are usually very important. A few words are given to indicate the sense in which each of the headings is intended, the headings themselves being kept very short in order to make them usable as mnemonics; some are expanded elsewhere in the Thesaurus.

The purpose of exhibiting the KEC here is partly to make the point that evaluation is an extremely complicated discipline, what one might call a multi-discipline. It cannot be seen as a straightforward application of standard methods in the traditional social science repertoire. In fact only seven of the eighteen checkpoints are seriously addressed in that traditional repertoire, and in most cases not very well addressed as far as evaluation needs are concerned.

1. *DESCRIPTION.* What is to be evaluated? The *evaluand*, described as objectively as possible. Does it have components? What are their relationships?

2. CLIENT. Who is commissioning the evaluation? The *client* for the evaluation; who may or may not be the *initiator* of the request for the evaluation; and may or may not be the *instigator* of the evaluand, e.g. its manufacturer or funding agency or legislative godparent; and may or may not be its *inventor* e.g. designer of a product or program.

3. *BACKGROUND & CONTEXT* of (a) the evaluand and (b) the evaluation. Includes identification of stakeholders (such as the non-clients listed in 2, the monitor, community representatives, etc.); believed nature of the evaluand; expectations from the evaluation; desired type of evaluation (formative vs. summative vs. ritualistic, holistic vs. analytical); reporting system; organization charts; prior efforts, etc.

4. RESOURCES ("Support System" or "Strengths Assessment") (a) available to or for use of the evaluand; (b) available to or for use of the evaluators. These are not what *is* used up, in e.g. purchase or maintenance, but what *could* be. They include money, expertise, past experience, technology, and flexibility considerations. These define the range of feasibility.

5. FUNCTION. What does the evaluand do? Distinguish what it is *supposed* to do—*intended or alleged* function or role—from what it *in fact* does—*actual* function(s) both for the client and the consumer; both *could* be covered under *Description* but it's usually best to treat them separately. Are there obvious dimensions or aspects or components of these functions?

6. DELIVERY SYSTEM. How does the evaluand reach the market? How is it maintained (serviced)? How improved (updated)? How are users trained? How is implementation achieved/monitored/improved? Who does all this?

7. *CONSUMER*. Who is using or receiving the (effects of the) evaluand? Distinguish *targeted* populations of consumers—*intended market*—from actually and potentially *directly* impacted populations of consumers—the *"true market"* or customers or recipients (or clients for the evaluand, often called the clientele); these should be distinguished from the total directly *or indirectly* impacted *recipient* population which makes up the *"true consumers."*

Note that the instigator, etc. (see 2 and 3) are also impacted, e.g. by having a job, but this does not make them consumers, in the usual sense. We should, however, consider them when looking at total effects and can describe them as part of the affected, impacted or involved group.

8. *NEEDS & VALUES* of the impacted and potentially impacted population. This will include wants as well as needs; and also values such as *judged* or believed standards of merit and ideals (cf. 9); the defined goals of the program where a goal-based evaluation is undertaken; and the needs etc. of the instigator, monitor, inventor etc., since they are indirectly impacted. The relative importance of these often conflicting considerations will depend upon ethical and functional considerations.

9. STANDARDS. Are there any pre-existing objectively validated standards of merit or worth that apply? Can any be inferred from CLIENT plus CONSUMER, FUNCTION and NEEDS/VALUES? (This will include *appropriate* ideals cf. the felt ideals in 8.) *If* goals are being considered, and *if* they can be validated as appropriate (e.g., from a needs assessment) *and* legal/ethical etc., they would graduate from being recorded in 8 to being accepted, as *one* relevant standard, in 9.

10. *PROCESS*. What constraints/costs/benefits apply to the normal *operation* of the evaluand (*not* to its effects or OUTCOMES (11))? In particular, legal/ethical-moral/political/managerial/aesthetic/hedonic/scientific? One *managerial* process constraint of special significance concerns the "degree of implementation," i.e., the extent to which the actual operation matches the program stipulations or sponsor's beliefs about its operation. One *scientific* process consideration would be the use of scientifically validated process indicators of eventual outcomes; another would be the use of scientifically (historically etc.) sound material in a textbook/course. One *ethical* issue would involve the relative weighting of the importance of meeting the needs of needy target population people and the career or status needs of other impacted-population people e.g. the program staff.

11. *OUTCOMES*. What effects are produced by the evaluand? (Intended or unintended). A matrix of effects is

useful; population affected × type of effect (cognitive/affective/psychomotor/health/social/environmental) × size of each × time of onset (immediate/end of "treatment"/later) × duration × each component or dimension (if analytical evaluation is required). For some purposes, the intended effects should be separated from the unintended (e.g. program monitoring, legal accountability); for others, the distinction should not be made (consumer-oriented summative product evaluation).

12. GENERALIZABILITY to other people/places/times/versions. ("People" means staff as well as recipients.) These can be labeled Deliverability and Saleability/Exportability/ Durability/ Modifiability.

13. *COSTS*. Dollar vs. Psychological vs. Personnel; Initial vs. Repeated (including Preparation-Maintenance-Improvement); Direct/Indirect vs. Immediate/Delayed-Discounted; by components if appropriate.

14. *COMPARISONS* with alternative options—include options recognized *and* unrecognized, those now available and those constructable—the leading contenders in this field are the "critical competitors" and are identified on cost plus effectiveness grounds. They normally include those that produce similar or better effects for less cost, and better effects for a manageable (RESOURCES) extra cost.

15. *SIGNIFICANCE*. A synthesis of all the above. The validation of the synthesizing procedure is often one of the most difficult tasks in evaluation. It cannot normally be left to the client who is usually ill-equipped by experience or objectivity to do it; and the formula approaches of e.g. cost-benefit calculations are only rarely adequate. "Flexible weighted-sum with overrides" is often useful.

16. RECOMMENDATIONS. These may or may not be requested, and may or may not follow from the evaluation; even if requested it may not be feasible to provide any, because the only type that would be appropriate are not such that any scientific evidence for specific ones is available in the relevant field of research. (RESOURCES available for the evaluation are crucial here.)

17. *REPORT*. Vocabulary, length, format, medium, time, location, and personnel for its (or their) presentation need careful scrutiny as does protection/privacy/publicity and

prior screening or circulation of final and preliminary drafts.

18. *METAEVALUATION.* The evaluation must be evaluated, preferably prior to (a) implementation, (b) final dissemination of report. External evaluation is desirable, but first the primary evaluator should apply the Key Evaluation Checklist to the evaluation itself. Results of the metaevaluation should be used formatively but may also be incorporated in the report or otherwise conveyed (summatively) to the client and other appropriate audiences. ("Audiences" emerge at metacheckpoint 7, since they are the "Market" and "Consumers" of the evaluation.)

KILL THE MESSENGER (phenomenon) The tendency to punish the bearer of bad tidings. One aspect of **valuephobia.** Much of the current attack on testing is pure KTM, like many of the elaborately rationalized earlier attacks on course grades. The presence of the rationalizations (in both cases) identify these as examples of a sub-species; Kill the Messenger—After a Fair Trial, of course.

LAISSEZ FAIRE (evaluation) "Let the facts speak for themselves." But do they? What do they say? Do they say the same thing to different listeners? Once in a while this approach is justified, but usually it's simply a cop-out, a refusal to do the hard professional task of synthesis and its justification. The laissez-faire approach is attractive to valuephobes—and to anyone else when the results are going to be controversial. The major risk in the naturalistic approach is sliding into laissez-faire evaluation, i.e.—to put it *slightly* tendentiously—no evaluation at all.

LEARNER VERIFICATION A phrase of Ken Komoski's, president of EPIE; refers to the process of (a) establishing that educational products actually work with the intended audience, and (b) systematically improving them in the light of the results of field tests. Now required by law in e.g. Florida and being considered for that status elsewhere. The first response of publishers was to submit letters from teachers testifying that the materials worked. This is not the R&D process that the term refers to. Some of the early

programmed texts were good examples of learner verification. Of course, it's costly, but so are four-color plates and glossy paper. It simply represents the application to educational products of the procedures of quality control and development without which other consumer goods are illegal or dysfunctional.

LEVEL OF EFFORT Level of effort is normally specified in terms of person-years of work, but on a small project might be specified in terms of person-months. It refers to the amount of direct "labor" that will be required, and it is presumed that the labor will be of the appropriate professional level; subsidiary help such as clerical and janitorial is either budgeted independently or regarded as part of the support cost, that is, included in a professional person-year of work. Person-years (originally man-years) is the normal unit for specifying level of effort. RFP's will often not describe the maximum sum in dollars that is countenanced for the proposal, but may instead specify it in terms of person-years. Various translations of a person-year unit into dollars are used; this will depend on the agency, the level of professionalism required, whether or not overhead and clerical support is separately specified, etc. Figures from $30,000 to over $50,000 per person-year are used at times.

LICENSING (of evaluators) See **Evaluation Registry**.

LITERARY CRITICISM The evaluation of works of literature; in many ways an illuminating model for evaluation—a good corrective for the emphases of the **social science model.** Various attempts have been made to "tighten up" literary criticism, of which the New Criticism movement is perhaps the best known, but they all involve rather blatant and unjustified preferences of their own (i.e., biases), exactly what they were alleged to avoid. The time is ripe to try again, using what we now know about **sensory evaluation**—and perhaps **responsive** and **illuminative** evaluation—to remind us of how to objectify the objectifiable while clarifying the essentially subjective. Conversely, a good deal can be learnt from a study of the efforts of F.R. Leavis (the doyen of the New Critics) and T.S. Eliot in his critical essays to precisify and objectify criticism. His view that "comparison and analysis are the chief tools of the critic" (Eliot, 1932), and even more his practice of displaying

very specific and carefully chosen passages to make points would find favor with the **responsive evaluators** today. Ezra Pound and Leavis went even further towards exhibiting the concrete instance (rather than the general principle) to make a point. This idiographic, anti-monothetic approach is not, contrary to much popular philosophy of science, anti-scientific as such; but in practice it failed to avoid various style or process biases, and too often (e.g. with Empson) became precious at the expense of logic. One can no more forget the logic of plot or the limits of possibility in fiction than the logic of function and the limits of responsibility in program evaluation.

LOCUS OF CONTROL Popular "affective" variable, referring roughly to the location someone feels is appropriate for the center of power in the universe on a scale from "inside me" to "far far away." A typical item might ask about the extent to which the subject feels s/he controls their own destiny. In fact, this is often a simple test of knowledge about reality and not affective (depending on how much stress is put on the feeling part of the item), and where it *is* affective, the affect may be judged as appropriate or inappropriate. So these items are usually misinterpreted, e.g. by taking any movement towards internalization of locus of control as a gain, whereas it may be a sign of loss of contact with reality.

LONGITUDINAL STUDY An investigation in which a particular individual or group of individuals is followed over a substantial period of time, in order to discover changes due to the influence of an evaluand or maturation, or environment. The contrast is with a **cross-sectional study**. Theoretically, a longitudinal study could also be an experimental study, but none of those done on the effect of smoking on lung cancer are of this kind although the results are almost as solid. In the human services area, it is very likely that longitudinal studies will be uncontrolled, certainly not experimentally controlled.

LONGTERM EFFECTS In many cases, it is important to examine the effects of the program or product after an extended period of time; often this is the *only* worthwhile criterion. Bureaucractic arrangements such as the difficulty of carrying funds over from one fiscal year to the next often

make investigation of these effects virtually impossible. "Longitudinal studies" where one group is "followed-up" over a long period are more commonly recognized as standard procedure in the medical and drug areas; an important example in education is the PROJECT TALENT study, now in its third decade. See **Overlearning.**

MAINTENANCE NEED A met but continuing **need.**

MAN-YEARS (properly, person-years) See **Level of Effort.**

MARKET The market checkpoint on the **Key Evaluation Checklist** refers to the disseminability of the product or program. Many needed products, especially educational ones, are unsaleable by available means. It is only possible to argue for developing such products if there is a special, preferably tested, plan for getting them used. No delivery system, no market.

MASSAGING (the data) Irreverent term for (mostly) legitimate synthesis of the raw results.

MASTERY LEVEL The level of performance actually *needed* on a criterion. Focus on mastery level training does not accept anything less, and does not care about anything more. Closely tied to **competency-based** approaches. Represents one application of **criterion-referenced testing.**

MATCHING See **Control Group.**

MATERIALS (evaluation) See **Product Evaluation.**

MATRIX SAMPLING If you want to evaluate a new approach to preventive health care (or science education), you do not have to give a complete spectrum of tests (perhaps a total of ten) to all those allegedly affected, or even to a sample of them; you can perfectly well give one or two tests to each in the sample, taking care that each test does get given to a random sub-sample, and preferably that it is randomly associated with each of the others, if they are administered pairwise (in order to reduce any bias due to interactions between tests). This will involve (a) much less cost to you than full testing of the whole sample, (b) less strain on each subject, (c) some contact with each, by con-

trast with giving all tests to a smaller sample, (d) ensuring that all of a larger pool of items get used on some students. The cost to each testee is much reduced, and the range of testees and items tested is much greater, both likely to be beneficial. But—the trade-off—you will not be able to say much about each individual. You are only evaluating the treatment's *overall* value. A good example of the importance of getting the evaluation question clear before doing a design.

MBO Management By Objectives, i.e. state what you're trying to do in language that will make it possible to tell whether you succeeded. Not bad as a guide to planning (though it tends to overrigidify the institution), but disastrous as a model for evaluation (though acceptable as *one* element in an evaluation design.) See **Goal-Based Evaluation.**

MEAN (Stat.) (Cf. **Median, Mode**) The mean score on a test is that obtained by adding all the scores and dividing by the number of people taking it; one of the several exact senses of "average." The mean is, however, heavily affected by the scores of the top and bottom few in the class, and can thus be non-representative of the majority.

MEASUREMENT Determination of the magnitude of a quantity, not necessarily, though typically, on a criterion-referenced test scale, e.g. feeler gauges, or on a continuous numerical scale. There are various types of measurement scale, in the loose sense, ranging from ordinal (grading or ranking) to cardinal (numerical scoring). The standard scientific use refers to the latter only. Whatever is used to do the measurement, apart—usually—from the experimenter, is called the instrument. It may be a questionnaire or a test or an eye or a piece of apparatus. In certain contexts, we treat the observer as the instrument needing **calibration** or validation. Measurement is a common and sometimes large component of *standardized* evaluations, but a very small part of its logic, i.e. of the justification for the evaluative conclusions.

MEDIAN (Stat.) (Cf. **Mean, Mode**) The median performance on a test is that score which divides the group into two, as nearly as possible; the "middle" performance. It

provides one exact sense for the ambiguous term "average." The median is not affected at all by the performance of the few students at the top and bottom of a class (cf. **Mean**). On the other hand, as with the mean, no one may score at or near the median, so that it doesn't identify a "most representative individual" in the way that the mode does. Scoring at the 50th **percentile** is (roughly) the same as having the median score, since 50% are below you and 50% above.

MEDIATED EVALUATION A more precise term for what is sometimes called (in a loose sense) process evaluation, meaning evaluation of something by looking at secondary indicators of merit, e.g. name of manufacturer, proportion of Ph.D.s on faculty, where someone went to college. The term "process evaluation" *also* refers to the *direct* check on e.g. ethicality of process.

MEDIATION (OR ARBITRATION) model of evaluation. Little attention has been paid to the interesting social role and skills of the mediator or arbitrator, which in several ways provides a model for the evaluator e.g. the combination of distancing with considerable dependence upon reaching agreement, the role of logic *and* persuasion, of ingenuity and empathy.

MEDICAL MODEL (of evaluation) In Sam Messick's version (in the ***Encyclopedia of Educational Evaluation***) the contrast is drawn between the engineering model and the medical model. The engineering model "focuses upon input-output differences, frequently in relation to cost." The medical model, on the other hand, (which Messick favors) provides a considerably more complex analysis, enough to justify: the treatment's generalization into other field settings; remediation suggestions; and side effect predictions. The problem here is that we cross the boundaries between evaluation and general causal investigations, thereby diluting the distinctive features of evaluation and so expanding its scope as to make results extremely difficult to obtain. It seems more sensible to appreciate ***Consumer Reports*** for what it gives us, rather than complain that it fails to give us explanations of the underlying mechanisms in the products and services that it rates. Cf. **Holistic** and **Analytic Evaluation**.

MERIT (Cf. **Worth**) "Intrinsic" value as opposed to extrinsic or system-based value/worth. For example, the merit of researchers lies in their skill and originality—their worth (to the institution that employs them) would include the income they generate.

META-ANALYSIS (Gene Glass) The name for a *particular approach* to synthesing studies on a common topic, involving the calculation of a special parameter for each ("Effect Size"). Its promise is to pick up something of value even from studies which do not meet the usual "minimum standards"; its danger is what is referred to in the computer programming field as the GIGO Principle—Garbage In, Garbage Out. While it is clear that a number of studies, none of which is statistically *significant*, can be integrated by a meta-analyst into a highly significant result (because the combined N is larger), it is not clear how *invalid* designs can be integrated. An excellent review of results and methods will be found in ***Evaluation in Education*** Volume 4, No. 1, 1980, a special issue entitled "Research Integration: the State of the Art". Meta-analysis is a special approach to what is called the general problem of research (studies) integration or research synthesis, and this array of terms for it reflects the fact that it is an intellectual activity that lies between data synthesis on the one hand and the evaluation of research on the other. As Light points out (ibid.) there is a residual element of **judgment** involved at several places in meta-analysis as in any research synthesis process; clarifying the basis for these judgments is a task for the evaluation methodologist and Glass' efforts to do so have led to the burgeoning of a very fruitful area of (meta-)research.

META-EVALUATION Meta-evaluation is the evaluation of evaluations, and hence typically involves using another evaluator to evaluate a proposed or completed evaluation. This practice puts the primary evaluator in a similar position to the evaluee; both are going to be evaluated on their performance. It can be done formatively or summatively. Reports should go to the original client, copy to the first-level evaluator for reaction. Meta-evaluation then gives the client independent evidence about the technical competence of the primary evaluator. No infinite regress is generated because extrapolation shows it doesn't pay after

the first meta-level on most projects and the second on any. Meta-evaluation is a professional obligation for evaluators, as psychoanalysis is for psychoanalysts. A dimensional approach might consider the validity, the credibility, utility (timeliness, readability, relevance), robustness and cost. The Key Evaluation Checklist can also be applied, in two ways: either by using it to generate a correct evaluation (or design), which can then be compared to the actual one (**secondary evauation**), or by applying the checklist *to* the original evaluation *as a product* (true metaevaluation). The latter process includes the former as an appropriate scientific process consideration; but it *also* requires us to look at e.g. the cost-effectiveness of the *evaluation* itself, and hence does something to assist the **balance of power**. It should, for example, normally include a look at the differential costs of Type 1 and Type 2 errors in the evaluation. Evaluations must *not*, however, be evaluated in terms of their actual consequences, but only in terms of their consequences *if used appropriately*. Besides the KEC, one might use the various **Evaluation Standards** or Bob Gowin's **QUEMAC** approach. Professionalism in evaluation requires regulation of the subject's self-reference and hence of the obligation to true meta-evaluation. See also **Consonance.**

MINIMUM COMPETENCY TESTING A basic level of (usually) basic skills is a minimum competency. Success in such tests has been tied to graduation, grade promotion, remedial education; failure has been tied to teacher evaluation, program non-funding etc. With all this at stake, MCT has been a very hot political issue—and an ethical one, and a measurement one. Introduced with due warning and support it can represent a step towards honest schooling; done carelessly, it is a disaster. See **Cutting Score.**

MISSION BUDGETING A generalization of the notion of program budgeting (see **PPBS**); the idea is to develop a system of budgeting which will answer questions of the type, "How much are we spending on such and such a *mission?*" (by contrast with *program*, *agency*, and *personnel*—the previous kinds of categories to which budget amounts were tied). One limitation of PPBS has been that a good many programs overlap in the clientele they serve and the services they deliver, so that we may have a very poor idea of how much we're putting into e.g. welfare or bilingual

education by merely looking at agency budgets or even PPBS figures, *unless* we have an extremely clear picture—which decision makers rarely can have, especially a new Executive Cabinet—of the actual impacted populations and the level of service delivery from each of the programs. This concept, along with **zero-based budgeting,** was popular with the early Carter administration but we hear little about it later in that regime, just as MacNamara's introduction of PPBS (into DOD, from Ford Motor Company) under an earlier administration has faded considerably.

MODE (Stat.) (Cf. **Mean, Median)** The mode is the "most popular" (most frequent) score (or score interval). It's more likely that a student about whom you know nothing except their membership in this group scored the "modal" score of the group than any other score. But it may not be *very* likely, e.g. if every student gets a different score, except two who get 100 out of 100, then the mode is 100, but it's not very "typical." In a "normal" curve, on the other hand, like the (alleged) distribution of IQ scores in the U.S. population, the mean, the median, and the mode are all the same value corresponding to the highest point of the curve. Some distributions, or curves representing them, are described as bi-modal, etc., which means that there are *two* (or more) peaks or modes; this is a looser sense of the term mode, but useful.

MODELS (of evaluation) A term loosely used to refer to a conception or approach or sometimes even a method (naturalistic, goal-free) of doing evaluation. Models are to paradigms as hypotheses are to theories, which means less general and some overlaps. Referenced here are the following, frequently referred to as models: **advocate-adversary, black box, connoisseurship, CIPP, discrepancy, engineering, judicial, medical, responsive, transactional** and **social science.** The best classification of these and others (many have been attempted) is Stufflebeam and Webster's (forthcoming, 1981).

MODUS OPERANDI METHOD A procedure for identifying the cause of a certain effect by detailed analysis of the chain of events preceding it and of the ambient conditions: it is sometimes feasible when a control group is impossible, and it is useful as a check or strengthening of

the design even when a control group *is* possible. The concept refers to the characteristic pattern of links in the causal chain which the criminalist refers to as the modus operandi of a criminal. These can be quantified and even configurally scored; the problem of identifying the cause can thus be converted into a pattern-recognition task for a computer. The strength of the approach is that it can be applied in individual cases, informally, semi-formally (as in criminalistics), and formally (full computerization). It also leads to MOM-oriented designs which deliberately employ "tracers" i.e. artefactual features of a treatment which will show up in the effects. An example would be the use of a particular sequence of items in a student questionnaire disseminated to faculty for instructional development use. (Details in a section by this title in ***Evaluation in Education,*** ed. W.J. Popham, McCutcheon, 1976.)

MONITOR The term "monitor" was the original term for what is now often called by an agency "the project officer," namely the person from the agency staff that is responsible for supervising progress and compliance on a particular contract or grant. "Monitor" was a much clearer term, since "project officer" could equally well refer to somebody whose responsibilities were to the project manager, or to somebody who merely handled the contract paper work (the "contract officer," as the fiscal agent at the agency is sometimes called). But it was apparently thought to have "Big Brother" connotations, or not to reflect adequately the full range of responsibilities, etc. See **Monitoring.**

MONITORING A **monitor** (of a project) is usually a representative of the funding agency who watches for proper use of funds, observes progress, provides information to the agency about the project and vice versa. Monitors badly need and rarely have evaluation skills; if they were all even semi-competent formative evaluators, their (at least quasi-) externality could make them extremely valuable since many projects either lack evaluation staff, or have none worth having, or never supplement them with external evaluation. Monitors have a schizophrenic role which few learn to handle; they have to represent and defend the agency to the project and represent and defend the project to the agency. Can these roles be further complicated by an

attempt at evaluation? They already include it and the only question is whether it should be done reasonably well.

MOTIVATION The disposition of an organism or institution to expend effort in a particular direction. It is best measured by a study of behavior, since self-reports are intrinsically and contextually likely to be unreliable. Cf. **Affect.**

MOTIVATIONAL EVALUATION The deliberate use of evaluation as a management tool to alter motivation can be content-dependent or content-independent. If the evaluation recommends a tie between raises and work-output which is adopted it may affect motivation; if it cuts the (supposed or actual) connection, it will be likely to have the opposite effect on motivation. But the mere announcement of an evaluation even without its occurrence, and certainly the presence of an evaluator, can have very large (good or bad) effects on motivation, as experienced managers well know. Evaluators, on the other hand, are prone to suppose that the contents of their reports are what counts, and tend to forget the reactive effects, while they would be the first to suspect the Hawthorne effect in a study done by someone else.

MULTIPLE-TIER See **Two-Tier.**

NATURALISTIC (evaluation or methodology). An approach which minimizes much of the paraphernalia of science e.g. technical jargon, prior technical knowledge, statistical inference, the effort to formulate general laws, the separation of the observer from the subject, the commitment to a single correct perspective, theoretical structures, causes, predictions and propositional knowledge. Instead there is a focus on the use of metaphor, analogy, informal (but valid) inference, vividness of description, reasons, explanations, inter-activeness, meanings, multiple (legitimate) perspectives, tacit knowledge. For an excellent discussion, see Appendix B: Naturalistic Evaluation in ***Evaluating with Validity***, E.House, Sage, 1980.) The Indiana University group (Guba and Wolf particularly) have paid particular attention to the naturalistic model and their definition (Wolf, personal communication) stresses: (a) more orienta-

tion towards "current and spontaneous activities, behaviors and expressions rather than to some statement of prestated formal objectives; (b) responds to educators, administrators, learners and the public's interest in different kinds of information; and (c) accounts for the different values and perspectives that exist . . ."; the approach stresses contextual factors, unstructured interviewing, observation rather than testing; meanings rather than mere behaviors. Much of the debate about the legitimacy/utility of the naturalistic approach recapitulates the idiographic/nomothetic debate in the methodology of psychology and the debates in the analytical philosophy of history over the role of laws. At this stage the principal exponents of the naturalistic approach (e.g. Stake) have gone too far in the laissez-faire direction (any interpretation the audience makes is allowable); but their example has shown up the impropriety of many of formalists' assumptions about the applicability of the **social science model.**

NEEDS ASSESSMENT (NEEDS SENSING is a related recent variant) This term has drifted from its literal meaning to a jargon status in which it refers to any study of the needs, wants, market preferences, values or ideals that might be relevant to e.g. a program. This sloppy sense might be called the "direction-finding" sense (or process), and it is in fact a perfectly legitimate process when one is looking for all possible guidance in planning or justification for continuance of a program. Needs assessment in the literal sense is just part of this *and it is the most important part*, hence, even if the direction-finding approach is taken, one must *then* sort out the true needs. Needs provide the first priority for response just because they are in some sense *necessary* whereas wants (merely) are *desired* and ideals are *"idealistic,"* i.e., often impractical. It is therefore very misleading to produce something as a NA (needs assessment) when in fact it is just a market survey because it suggests that there is a level of urgency or importance about its findings which simply isn't there. True needs are considerably harder to establish than felt wants, because true needs are often unknown to those who have them—possibly even contrary to what they want, as in the case of a boy who needs a certain diet and wants an entirely different one.

The most widely used definition of need—the "dis-

crepancy definition"—does not confuse needs with wants but does confuse them with ideals. It defines need as the gap between the actual and the ideal, or whatever is needed to bridge it. This definition has even been built into law in some states. But the gap between your actual income and your ideal income is quite different (and much larger) than the gap between your actual income and what you absolutely need. So we have to drop the use of the ideal level as the key reference level in the definition of need—which is just as well, because it is very difficult to get much agreement on what the ideal curriculum is like and if we had to do that before we could argue for any curriculum needs, it would be hard to get started.

A second fatal flaw in the discrepancy definition is its fallacious identification of needs with one particular subset of needs, namely *unmet* needs. But there are many things we absolutely need—like oxygen in the air, or vitamins in our diet—which are already there. To say we need them is to say they are *necessary* for e.g. life or health, which distinguishes them from the many *inessential* things in the environment. Of course, on the discrepancy definition they are not needs at all, because they are part of "the actual," not part of the gap (discrepancy) between that and the ideal. It may be useful to use the dietary terminology for met and unmet needs—*maintenance* and *incremental* needs. People sometimes think that it's better to focus on incremental needs because that's where the action is required (so maybe the discrepancy definition doesn't get us into too much trouble). But where will you get the resources for the necessary action? Some of them usually come from redistribution of existing resources, i.e., from robbing Peter's needs to pay for Paul's, where Peter's (the maintenance needs) are just as vital as Paul's (the incremental). This leads to an absurd flip-flop in successive years: it is much better to look at all needs in the NA, prioritize them (using **apportioning** methods *not* **grading** or **ranking**) and then act to redistribute old and new resources.

The correct definition of need, which we might call the *diagnostic definition*, defines need as anything essential for a satisfactory mode of existence, i.e., anything without which that mode of existence or "level of performance" would fall below a satisfactory level. The slippery term in this is of course "satisfactory" and it is context-dependent; satisfac-

tory diets in a nation gripped by famine may be considerably nearer the starvation level than those regarded as satisfactory in a time of plenty. But that is part of the essentially pragmatic component in NA—it is a prioritizing and pragmatic concept. Needs slide along the middle range of the spectrum from disaster to utopia as resources become available. They never cover the ends of the spectrum—no riches, however great, legitimate the claim that everyone needs all possible luxuries.

The next major ambiguity or trap in the concept of need relates to the distinction between what we can call *performance needs* and *treatment needs*. When we say that children need to be able to read, we are talking about a needed level of performance. When we say they need classes in reading, or instruction in the phonics approach to reading, we are talking about a needed treatment. The gap between the two is vast, and can only be bridged by an evaluation of the alternative possible treatments that could yield the allegedly needed performance. Children need to be able to converse—but it does not follow they need classes in talking, since they pick it up without any. Even if it can be shown that they do need the "treatment" of reading classes, that's a long way from the conclusion that any particular approach to reading instruction is needed. The essential points are that the kind of NA with which one should begin evaluations is *performance* NA; and that *treatment* needs claims essentially require *both* a performance NA *and* a full-scale evaluation of the relative merits of the best candidates in the treatment stakes.

Conceptual problems not discussed here include the problem of whether there are needs for what isn't feasible, and the distinction between artificial needs (alcohol) and essential needs (food); methodological problems including the flaws in the usual procedures for performing NA are discussed elsewhere (***LE***).

The crucial perspective to retain on NA is that it is a process for discovering *facts* about organisms or systems; it's not an opinion survey or a wishing trip. It is a fact about children, in this environment, that they need Vitamin C and functional literacy skills, whether or not they think so or their parents think so or for that matter witchdoctors or nutritionists or reading specialists think so. What makes it a

fact is that the withdrawal of, or failure to provide these things, results in very bad consequences, by any reasonable standards of good or bad. Thus, models for NA must be models for truth-finding, not for achieving political agreement. That they are all too often of the latter kind reflects the tendency of those who design them to think that value judgments are not part of the domain of truth. For NA *are* value judgments just as surely as they are matters of fact; indeed, they are the key value judgments in evaluation, the root source of the value that eventually makes the conclusion an evaluative one rather than a purely descriptive one. It's easy to see this if we began with a statement that referred to an ideal as we (implicitly) do with the discrepancy definition; or if we had a treatment-need statement to *start* off (since that *is* an evaluation). And it's easy to see that if we began with mere market surveys, we would *not* have an evaluative conclusion, just a descriptive one (possibly describing a population's evaluation, but not *making* evaluations). But diagnostic-definition performance NAs *are* evaluative because they require the identification of the *essential*, the *important*, that which avoids *bad* results. Of course, these are often relatively uncontroversial value judgments. Evaluations build on NAs like theories build on observations; it's not that observations are infallible, only that they're much *less* fallible than theoretical speculation.

NORMAL DISTRIBUTION (Stat.) Not the way things are normally distributed, though some are, but an ideal distribution which results in the familiar bell-shaped curve (which, for example, is perfectly symmetrical though few real distributions are). A large part of inferential statistics rests on the assumption that the population from which we are sampling is normally distributed, with regard to the variables of interest, and is invalid if this assumption is grossly violated as it quite often is. Height and eye color are often given as examples of variables that are normally distributed but neither are well-supported examples. (The term "Gaussian distribution" is sometimes and much less confusingly used for this distribution.)

NORM-REFERENCED TESTS These are constructed to yield a measure of *relative* performance of the individual (or group) by comparison with the performance of other

individuals (or groups) taking the same test e.g. in terms of percentile ranking (cf. **Criterion-Referenced Tests**). Since the simplest and often the best quick way to determine whether a test involves unrealistic standards is by finding out how many students in the state succeed, at that level norm-referencing is a valuable part of any testing program. It is not ideal as a *sole* basis since it makes discriminating or competing more important than (or the only meaning of) achieving, and severely weakens the test as an indicator of mastery (or excellence or weakness), which you should also know about. The best compromise is a criterion-referenced test on which the norms are also provided, whose criteria are documented needs.

NULL HYPOTHESIS The hypothesis that results are due to chance. Statistics only tells us about the null hypothesis; it is experimental design that provides the basis for inferences to the truth of the scientific hypothesis of interest. The "significance levels" referred to in experimental design and interpretation are the chances that the null hypothesis is correct. Hence, when results "reach the .01 level of significance" that means there's only one chance in a hundred that they would be due to chance. It does *not* mean that there's a 99 percent chance that *our* hypothesis is correct; because, of course, there may be other explanations of the result that we haven't thought of.

NUT ("making the nut") Management consultant jargon for the basic cost of running the business for the year. After "making the nut" one may become a little choosier about which jobs to take on, and what rates to set.

OBJECTIVES The technical sense of this term refers to a rather specific description of an intended outcome; the more general description is referred to as a goal.

OBSERVATION The process or product of direct sensory inspection, frequently involving trained observers. The line between observation and its normal antonym "interpretation" is not sharp and is in any case context-dependent, i.e. what counts as an observation in one context ("a very pretty dive") will count as an interpretation in another (where the diving judges' score is appealed). Just as

it is very difficult to get trainees in evaluation—even those with considerable scientific training—to write non-evaluative descriptions of something that is to be evaluated, so it is difficult to get observers to see only what's there rather than their inferences from it. The use of checklists and training can produce very great increases in reliability and validity in observers; observation is thus a rather sophisticated process, and not to be equated with the amateur's perceptions or reports on them. It should be clear from the above that there are contexts in which observers, especially trained observers, can correctly report their observations in evaluative terms. (An obvious example, where no special training is involved, is reporting scores at a rifle range.)

OPPORTUNITY COSTS Opportunity costs are what one gives up by engaging in a particular activity. The same concept applies to investing money or any other resource. There are *always* opportunity costs; one at least has to give up leisure to do something, or give up work to do nothing, i.e., enjoy leisure. Calculating them (like **profit**) is a conceptual task first, and an arithmetic one later. In the first place, there is always an infinity of alternatives to any action, all of which one gives up. Does it follow that opportunity costs are always infinite? The convention is that the OP is the value of the *most valuable* of these. So, calculating one OP often involves calculating a gret many costsof alternatives.

OUTCOME EVALUATION See **Pay-off Evaluation.**

OVERLEARNING Overlearning is learning past the point of 100% recall, and is aimed at generating long-term retention. In order to avoid boredorn on the part of the learner, and for other reasons, the best way to do this is through reintroducing the concept (etc.) in a variety of different contexts. One reason that long-term studies, or the follow-up phase of an evaluation often reveals grave deterioration of learning is that people have forgotten the distinction between learning to criterion at t_1 and learning to criterion at t_2; in fact, the latter *is* the correct criterion, where t_2 is the time when the knowledge is needed, while t_1 is the end of the instructional period.

PAD, PADDING When a bidder goes up with a budget for a proposal, there has to be—one way or another—some allowance in it for unforeseen eventualities—at least if it is to be done according to sound business practices. This is often referred to as "the pad," and the practice of doing this is the *legitimate* version of "padding the budget." Padding the budget is also used as a term to refer to illegitimate additions to the budget (excessive profits); but it must be realized that the pad is the only recourse that the contractor has for handling the obvious unreliabilities in predicting the ease of implementing some complicated testing program, the ease of designing a questionnaire that will get past the questionnaire review panels, etc.

PARADIGM An extremely general conception of a discipline, which may be very influential in shaping it, e.g. "the classical social science paradigm in evaluation."

PARALLEL FORMS Versions of a test that have been tested for equal difficulty and validity.

PARALLEL PANELS In proposal review, for example, it is important to run independent concurrent panels in order to get some idea of the reliability of the ratings they are producing. On the few occasions this has been done, the results have been extremely disquieting. Unreliability guarantees both invalidity and injustifce. On would expect a federal science foundation to have enough commitment to validity and justice to do routine checks of this kind, they usually cry poormouth instead of looking for ways to get validity within the same budget. In any case, dispensing funds invalidly and unfairly is not justified by saying it would cost slightly more to do it reasonably well, even if true, since the payoffs would be higher (from the definition of "doing it, and reasonably well"), and justice is supposed to be worth a little.

PARETO OPTIMAL A tough criterion for changes in e.g. an organization or program which requires that changes be made only if nobody suffers and somebody benefits. Crucial feature is that it appears to avoid the problem of justifying so-called "interpersonal comparisons of utility," i.e., showing that the losses some sustain as a result of a change are less important than the gains made by others. Improving welfare conditions by raising taxes is *not*

pareto optimal, obviously. But selecting between alternative pareto optimal changes *still* involves relative hardship and benefit considerations. A major weakness in Rawl's theory of justice is the commitment to Pareto optimality.

PARETO PRINCIPLE A management maxim possibly more illuminating than the **Peter Principle** and **Parkinson's Law;** it is sometimes described as the 80/20 rule, or the "principle of the vital few and the trivial many," and asserts that about 80% of significant achievement e.g. at a meeting is done by about 20% of those present; 80% of the sales come from 20% of the salespeople, 80% of the pay-off from a task-list can be achieved from 20% of the tasks, etc. Worth remembering because it's sometimes true, and often surprising.

PARKINSON'S LAW "Work (and budgets, timelines and staff size), expands to fill the space, time and funds available." If its converse were only true it would mean we could do everything by allowing no time for it; but *it* is an insight about large organizations. The fact that bids on RFP's come in close to the estimated limit may not illustrate this, but only that the work could be done at various levels of thoroughness, or that RFP writers aren't dumb.

PASSIVE (evaluation) See **Active.**

PAYBACK (PERIOD) A term from fiscal evaluation which refers to the time before the initial cost is recovered; the recovery cash flows should of course be **time-discounted.** Payback analysis is what shows that buying a $12,000 word-processor may be sensible even if the price will probably drop to $8000 in a year; if the payback period is, say, 15 months (typical of many carefully-chosen installations), you will in fact lose several thousand dollars by waiting in the belief that.

PAYOFF EVALUATION Evaluation focused on results; the method of choice apart from costs, delay, and intervening loss of control or responsibility (See **Process Evaluation.**) Essentially similar to outcome evaluation.

PERCENTILE (Stat.) If you arrange a large group in the order of their scores on a test, and divide them into 100 equal-sized groups, beginning with those who have the lowest score, the first such group is said to consist of those

in the 1st percentile (i.e., they have scores worse than 99% of the group), and so on to the top group which should be called the 100th percentile: for boring technical reasons the actual procedure used only distinguishes 99 groups, so the best one can do is get into the 99th percentile. With smaller numbers or for cruder estimates, the total group is divided into ten *deciles;* similarly for four *quartiles,* etc.

PERFECTIONISM Marks' Principle: "The price of perfection is prohibitive." Never get letters or papers retyped when fully legible corrections can be made by hand; there aren't enough trees, days or dollars for that. Legal documents and typographical works of art may be exceptions, but the Declaration of Independence has two insertions by the scribe so there's a precedent in a legal case. (Cited by Bliss.

PERFORMANCE CONTRACTING The system of hiring and paying someone to deliver (e.g. educational) services by results. They might be paid in terms of the number of students times the number of grade equivalents their scores are raised. Widely tried in the 60s, now rare. Usual story is that it didn't work or worked only by the contractor's staff cheating ("teaching to the test"). Actual situation was that the best contractors did a consistently good job but the *pooled* results of all contractors were not good. As with most innovations, the total lack of sophistication (in evaluation) of the educational decision makers treated this result as grounds for giving up, instead of for hiring the better contractors, from which we might have gone on to still better teaching methods for everyone. See **regression to the mean** for an example of the need for some sophistication in setting the terms of the contract.

PERSONNEL EVALUATION Personnel evaluation typically involves an assessment of job-related skills, in one or more of five ways; first, judgmental observation of job-performance by untrained but well-situated observers e.g. coworkers; second, judgmental observation by skilled observers e.g. experienced supervisor or personnel manager or consultant; third, direct measurement of job performance parameters, by calibrated instruments (human or, usually, other); fourth, observation or measurement of performance on job simulations; fifth, the same on paper and pencil tests

which examine job-relevant knowledge or attitudes. Personnel evaluation not only involves ethical constraints upon the way it should be done, it must also involve an ethical dimension on which the performance of the personnel is scored. The importance of that will vary depending upon the amount of authority and interpersonal contact of the individual being evaluated. There are a number of standard traps in personnel evaluation which invalidate most of the common approaches. For example, the failure to provide appropriate levels of anonymity for the raters, consistent with relevant legislation, *or* a general fear of badmouthing others because it involves the sin referred to in "judge not that ye be not judged," leads to an unwillingness to voice criticism even if deserved; this (solvable) problem requires sustained and ingenious attention. The scales used in personnel evaluation are rarely based upon serious job analysis and consequently can hardly give an accurate picture of someone's performance. Another common mistake is to put style variables into evaluation forms or reports, in situations where no satisfactory evidence exists that a particular style is superior to others. Even when style variables have been validated as indicators of superior performance, they typically cannot be used in personnel evaluation because the correlations between their presence and good performance are merely statistical, and are thus as illegitimate in the evaluation of *individuals* as skin color, which of course does correlate statistically with various desirable and/or undesirable characteristics. "Guilt by association" is as inappropriate when the association is via a common style as when it is via a common friend, race or religion.

PERSON-YEARS See **Level of Effort.**

PERSPECTIVAL EVALUATION This approach to or part of an evaluation requires the evaluator to attempt various conceptualizations of the program or product being evaluated. Programs and products can be seen from many different perspectives which affect every aspect of the evaluation, including cost analysis. Advocate-adversary is a special case of perspectival evaluation; consumer-based or manager-based evaluations are special perspectives. As in architecture, multiple perspectives are required in order to see something in full depth. Different from **illuminative,**

responsive etc. in the total commitment to the view that there *is* an objective reality *of which* the perspectives are merely views and *inaccurate* by themselves. The correct strand in the *naturalistic* approach stresses this; the weak strand favors the "each perspective is legitimate" approach, which is false if the perspective is claimed to be *the* reality and not *just* one aspect of it.

PERT, PERTCHART Stands for **P**rogram **E**valuation **R**eview **T**echnique; a special type of **flow chart**ing, of which perhaps the most interesting feature is the fact that an effort is made to project times at which various points in the project's developent will be reached (and outputs at those points) at three levels, namely the maximum likely, the minimum likely and the most probable (date or level). This provides a good approach to contingency planning, in the hands of a skilled manager. As with all these devices, they can become a pointless exercise if not closely tied to reality, and the tie to reality can't be read off the chart.

PLACEBO EFFECT The effect due to the *delivery context* of a treatment as opposed to the *delivered content*. In medicine, the placebo is a dummy pill, given to the control group in exactly the same way as the test drug (or more generally, the experimental treatment) is given to the experimental group, i.e., with the nurses, doctors and patients in ignorance as to whether the pill is a placebo or not. (Notice that there are two errors in this as a valid design for identifying placebo effect, but it's a considerable improvement over giving no placebo to the control group.) "Bedside manner" carries the placebo effect with it and since it is estimated that prior to the sulfa drugs, 90% of all therapeutic results were due to the placebo effect, it's a little unfortunate that bedside manner gets little play in medical practice and training (and, until 1948, no research). Psychotherapy has been said to be *entirely* placebo effect (Frank); a design to investigate this view presents interesting challenges. In education and other human service areas, the placebo effect is roughly equivalent to the Hawthorne effect which probably accounts for most successes with innovations. This is as licit as bedside manner, but only if not ascribed to the snake-oil itself. But if we're honest about it being only a placebo, won't the placebo effect evaporate? Not if the charismatic context is preserved; "the heart has its reasons that Reason

doesn't know."

PLANNING (evaluation in) See **Preformative Evaluation.**

POINT CONSTANCY REQUIREMENT (PCR) The requirement on numerical scoring e.g. of tests that a point, however earned (i.e. on whatever item and for whatever increment of performance on a particular item), should reflect the same amount of merit. (It is connected with the definition of an interval scale.) If the PCR is violated, additivity fails i.e. performance A will add up to more points than performance B although it is in fact inferior. PCR is a very severe requirement and rarely even tried for in any serious way, hence one should normally (holistically) grade as well as score tests to provide protection against PCR failure. The key to PCR is the **rubric** in essay/simulation scoring and item-tailoring on multiple-choice tests.

POINT OF ENTRY The point of entry problem is the problem for the *client* of when to bring an evaluator in on a project, and the problem for the *evaluator* of the point in the time flux of decisions when s/he should start evaluating the options (critical competitors). Project directors and program managers often feel that bringing in an external evaluator (and often any evaluator at all) at the very beginning of a project is likely to product a "chilling effect," and that the staff should have a chance to "run with the ball" in the way they think is most likely to be productive for at least some time without admonitions about measurability of results, etc. The result is often that the evaluator is brought in too late to be able to determine base-line performance, and too late to set up control groups and is hence unable to determine either gains or causation, to mention only two of the major problems that occur in trying to do evaluations of projects that were designed without evaluation in mind. This is not to say that evaluators never or rarely exert a chilling effect; they often do. Often they could have avoided it; sometimes not. (**GFE** is one way to avoid it but impossible in the planning phase.) It's possible on a small project to have an evaluator in for at least one *series of discussions* during the planning phase, maybe get by without one for a while after that, bring one or more back in after things begin to take shape, and perhaps dispense with most of them

again for a second period of "unfettered creativity." However, there are many good evaluators that exert a constantly supportive and helpful effect on projects, in spite of being on board all the time. They will need external evaluation help to avoid the bias of co-option, but on a *big* project there's really no alternative to an in-house early-on-board evaluation staff. From the evaluator's point of view, the question is what to consider "fixed," what to consider as beyond second-guessing in doing an evaluation. Suppose that one is brought in very late in a project. For formative evaluation purposes, there's really no point in second-guessing the early decisions about the form of the project, because they're presumably irreversible. For summative evaluation, it *will* be necessary to second-guess those, and that means that the point of entry of the summative evaluator will be back at the moment when the project design was being determined, a point which presumably antedates the allocation of funds to the project. The formative evaluator, however, should in fact not be restricted to looking at the set of choice points that are seen by the project staff as downstream from the point at which the evaluator is called in. For the formative evaluator, the correct point of entry for evaluation purposes is the *last irreversible decision*. Even though the staff hasn't thought of the possibility of reversing some earlier decisions, the formative evaluator must look into such possibilities and the cost/value of reversals.

POLITICS OF EVALUATION Depending on one's role and the day of the week, one is likely to think of politics as *dirty* politics—an intrusion into scientific evaluation—or as part of the ambient *reality* which evaluators are too often too careless about including as relevant considerations. If one has a favorable attitude towards politics, or uses the term without pejorative connotations, one will include virtually all program background and contextual factors in the political dimension of program evaluation. The jaundiced view simply defines it as the set of pressures that are not related to the truth or merits of the case. The politics of competency-based testing as a requirement for graduating is a good example. The situation in many states is that it has become "politically necessary" to institute such requirements, now or in the near future, although the *way* in which they have been instituted virtually destroys all the reasons

for the requirements. That is, the requirement for graduating from the 12th grade is "basic skills" at the 7th or 8th grade level; no demonstration of *other* skills; not even any demonstration of *application* skills on the basics; the exams set up so that multiple retakes of exactly the *same* test are possible (hence there is no proof that the *skill* is present); teachers have access to and teach to the test; other subjects are completely dropped from the 11th and 12th grade curriculum in order to make room for yet more repetitive teaching of drill-level basics, etc. A strong case can be made that *this* version of MCT does more harm than good, though a genuine version would certainly contribute towards truth-in-packaging of the diplomate. This is politics without payoff. But on many occasions, the "politics" is what gets equity into personnel evaluation, and racism out of the curriculum, though it also keeps moral education out of the public schools, a terrible handicap for the society. Better education about and in evaluation is the only hope of improvement, short of a political leader with the charisma to persuade us of anything and the brains to persuade us to improve our self-critical skills.

POPULATION (Stat.) The group of entities from which a sample is drawn, or about which a conclusion is inferred. Originally meant people, obvious extension to things (e.g. objects on the production line, the population which is sampled for quality control studies); less obvious extensions to circumstances (a field trial samples the population of circumstances under which a product might be used); still fancier extensions in statistical theory to possible configurations, etc.

PORTRAYAL Semi-technical term for an evaluation-by-(rich)-description, perhaps using pictures, quotes, anecdotes as well as observations. See **Responsive Evaluation, Naturalistic Evaluation.**

POSTTEST The measurement made after the "treatment," to get absolute or relative gains (depending on whether the comparison is with pre-test scores or comparison group scores.)

POWER (of a test, design, analysis) An important technical concept involved in the evaluation of experimental designs and methods of statistical analysis, related to effi-

ciency. It is in tension with other desiderata such as small sample size, as is usual with evaluative criteria.

PPBS Program Planning and Budgeting System. The management tool developed by MacNamara and others at Ford Motor Company and taken to the Pentagon when MacNamara became Secretary of Defense; since then widely adopted in other federal and state agencies. Principal advantage and feature: identifying costs by *program* and not by conventional categories such as payroll, inventory, etc. Facilitates rational planning with regard to program continuance, increased support, etc. Two problems: first, it's too often (virtually always) instituted as a mere change in bookkeeping procedures, without a program *evaluation* component worth the name, so the gains in decision validity don't occur. Second, it's often very expensive to implement and unreliable in distribution of overhead and it never seems to occur to anyone to evaluate the problem and cost of shifting to PPBS before doing it, a typical example of missing the point of the whole enterprise. Cf. **Meta-evaluation, Mission Budgeting.**

PRACTICE EFFECT The *specific* form of practice effect refers to the fact that taking a second test with the same or closely similar items in it, will result in improvement in performance even if no additional instruction or learning has occurred between the two tests. After all, one has done all the "organizing of one's thoughts" before the second test. There is a *general* practice effect, which is particularly important with respect to individuals who have not had much recent experience with test-taking; this practice effect simply refers to improving one's test-taking skills through practice, e.g. one's ability to control the time spent on each question, to understand the way in which various types of multiple-choice questions work, etc. The more **speeded** the test is, the more serious the practice effect is likely to be. The use of control groups will enable one to estimate the size of the practice effect, but where they're not possible, the use of a posttest-only design for *some* of the experimental group will do very nicely instead, since the difference between the two sub-groups on the posttest will give an indication of the practice effect, which one then subtracts from the gains of the posttest-only group in order to get a measure of the gains due to the treatment.

PREDICTIVE VALIDITY See **Construct Validity**

PREFORMATIVE (evaluation) Evaluation in the planning phase of a program; typically involves gathering baseline data, improving evaluability, designing the evaluation, improving the planned program etc. See **Evaluation.**

PRESS RELEASES The rules are: (1) Don't bother to hand (or send) out the technical version, even as a supplement. (2) Don't bother to hand out a summary of the technical document. (3) Don't bother to hand out a statement which says favorable things and then qualifies them—either the qualification or the favorable comment will be dropped. (4) Issue only a basic description of the program itself plus a single overview claim, e.g. "Results do not as yet show any advantages or disadvantages from this approach, because it's much too early to tell. May have a definite conclusion in n months." (That's an interim release; in a final release you drop the second sentence.)

PRETEST Pretests are normally said to be of two kinds; diagnostic and baseline. In a diagnostic pretest, the pedagogical (health etc.) function is to identify the presence or absence of the prerequisite skills, or the places where remediation instruction should be provided. These tests will typically not be like the posttests. In baseline pretesting, on the other hand, we are trying to determine what the level of knowledge (etc.) is on the criterion or pay-off dimensions, and hence it should be matched exactly, for difficulty, with the posttest. Instructors often think that using this kind of pretest will have bad results, because students will have a "failure experience". Properly managed, the reverse is the case; not only does one frequently discover that some or all students are not as ignorant as one had thought about the subject matter of the course, in which case very useful changes can be made in content, or "challenging out" can be allowed, with a reduction in costs to the student and possibly to the instructor. Moreover, the pretest gives an excellent and highly desirable preview of the kind of work that will be expected, and if it is—as it should be—gone over carefully in class, one has provided students with an operational definition of the required standards for passing. Furthermore, one has created a quite useful climate for interesting the students in early discussions, by giving them

a chance to try to solve the problems with their native wit, and then explaining how the content of the course helps them to do better. In many subjects, though not all, this constitutes a very desirable proof of the importance of the course. Of course, treating the pretest as defining the early course content is likely to qualify as **teaching to the test** if one uses many of the items from the pretest in the posttest. But there are times when this is entirely appropriate; and in general it is very sensible to pull items for the posttest out of a pool that includes the items from the pretest, so that at least some of them will be retested. This encourages learning the material covered in the pretest, which should certainly *not* be excluded from the course just because it has already been tested. Instructors who begin to give pretests also begin to adjust their teaching in a more flexible way to the requirements of a specific class, instead of using exactly the same material repeatedly. Thus the use of a pretest is an excellent example of the integration of evaluation into teaching, *and* a case of evaluation procedures paying off through side effects as well as through direct effects (which, in this case, would be the discovery that students are not able to learn certain types of material from the text, notes and lectures provided on that topic.)

PROCESS EVALUATION Usually refers to evaluation of the treatment (or evaluand) by looking at it instead of the outcome. With exceptions to be mentioned, this is only legitimate if some connection is *known* (not *believed*) between process variables and outcome variables, and it is never the best approach because such connections, where they do exist, are relatively weak, transient, and likely to be irrelevant to many new cases. The classic case is evaluation of teachers by classroom observation (the universal procedure K–12), where there are *no* evaluation-useable connections between classroom behavior and learning outcomes, quite apart from the problem that the observer's presence produces atypical teaching behavior, and the observer is normally someone with other personal relations with the teacher that are highly conducive to bias. (The evaluation of administrators is no better.) Certain aspects of process *should* be looked at, as *part* of an evaluation, not as a substitute for inspection of outcomes, e.g. its legality, its morality, its enjoyability, implementation of alleged treatment, and

whether it can provide any clues about causation. It is better to use the term **mediated evaluation** to refer to what is described in the opening sentence of this entry, and allow process evaluation to refer to that *and* to the direct evaluation of process variables as part of an overall evaluation which involves looking at outcomes.

PRODUCT Interpreted very broadly, e.g. may be used to refer to students, etc., as the "product" of a training program; a pedagogical *process* might be the *product* of a Research and Development effort.

PRODUCT EVALUATION The best-developed kind of evaluation; ***Consumer Reports*** used to be the paradigm though it has deteriorated significantly in recent years (***PE***). See **Key Evaluation Checklist.**

PROFESSIONALISM, PROFESSIONALITY Somewhere above minimum competence in a profession but short of the realm of professional ethics there is a set of obligations e.g. to keeping current, and to self-evaluation, which should be supported and counted in personnel evaluation. Professional ethics for quarterbacks prohibits kick-backs, professionalism requires kicking practice.

PROFIT This term from fiscal evaluation has unfortunate connotations to the uninformed. The gravity of the misconception becomes clear when a non-profit organization starts doing serious budgeting and discovers that it has to introduce something which it can scarcely call profit, but which does the same job of funding a prudent reserve, new programs and buildings, etc. (It calls it "contribution to margin," instead.) The task of defining profit is essentially a philosophical one. Granted that we should distinguish gross profit from net profit and that gross "profit" has to cover all overhead (e.g. administrative, amortization, insurance and space expenses) which may leave no (net) profit at all, what should we do about the cost of the money capital and time invested when both are furnished by a proprietor/manager or by donors? Is a proprietor whose "net profit" covers his time at the rate of $5 per hour really making a profit? If ROI on the capital investment is 3% in a market which pays 10% on certificates of deposit, is this "making a profit" or a loss when s/he could make $20/hour in salary?

Using **opportunity cost** analysis, the answer is, No; but the usual analysis says, Yes. That's correct for the Internal Revenue Service, but not for employees considering a strike. As usual, **cost analysis** turns out to be conceptually very complex although few people realize this; consequently serious mistakes are very common. If the buildings (or equipment) have been amortized completely, should one deduct a slice of the eventually-necessary replacement cost down-payment before one has a profit? Should *some* recompense for risk (or prior losses) be allowed before we get to "profits"? Cost analysis/fiscal evaluation looks precise because it's quantitative, like statistics, but eventually the conceptual/practical problems have to be faced and most current definitions will give you absurd consequences, e.g. "the business is profitable, but I can't afford to keep it going."

PROGRAM The general effort which marshals staff and *projects* towards some (often poorly) defined and funded goals.

PROGRAMMED TEXT One in which the material is broken down into small components ("frames"), ranging in length from one sentence to several paragraphs, within which some questions are asked about the material, e.g. by leaving a blank which the reader has to fill in with the correct word, possibly from a set of options provided. This *interactive* feature was widely proclaimed to have great virtue in itself. It had none, *unless* very thorough R&D effort was also employed in the process of formulating the exact content and sequence of the frames and choices provided. Since the typographic format does not reveal the extent of the field-testing and rewriting (and hence conceals the total absence of it), lousy programmed texts quickly swamped the market (late 50s) and showed that Gresham's Law is not dead. As usual, the consumers were mostly too naive to require performance data and the general conclusion was that programmed texts were "just another fad." In fact, the best ones were extremely powerful teaching tools, *were* in fact "teacher-proof" (a phrase which did not endear them to one group of consumers), and some are still doing well (Sullivan/BRL reading materials, for example). A valiant effort was made by a committee under Art Lumsdaine to set up standards, but the failure of all professional training programs to teach their graduates serious evaluation skills

meant there was no audience for the standards. We shall see whether the new Evaluation Standards from the Stufflebeam group suffer a better fate.

PROJECTS Projects are time-bounded efforts, often of a program.

PROJECTIVE TESTS These are tests with no right answer; the Rorschach inkblot test is a classic example, where the subject is asked to say what s/he sees in the inkblot. The idea behind projective tests was that they would be useful diagnostic tools, and it seems quite possible that there are clinicians who do make good diagnoses from projective tests. However, the literature on the **validity** of Rorschach interpretations, i.e. those which can be expressed verbally as unambiguous rules for interpretations, is essentially negative. The same is unfortunately true of many other projective tests, which fail to show even test-retest reliability, let alone interjudge reliability (assuming that shared bias is ruled out by the experimental design), let alone predictive validity. Of course, they're a lot of fun, and very attractive to valuephobes—both testers and testees—since there are no right answers.

PROTOCOL See **Evaluation Etiquette.**

PSEUDO-NEGATIVE EFFECT An outcome or datum that appears to show that an evaluand is having exactly the wrong kind of effect, whereas in fact it is not. Four paradigm examples are: the Suicide Prevention Bureau whose creation is immediately followed by an increase in the rate of reported suicides; the school intercultural program which results in a sharp rise of interracial violence; the college faculty teaching improvement service whose clients score worse than non-clients; the drug education (or sex education) program which leads to "experimentation." (See text of ***Introduction to Evaluation***, Scriven. for treatment of these examples.)

PSEUDO-POSITIVE EFFECT Typically, an outcome which is consistent with the goals of the program, but in circumstances where *either* the goals *or* this way of achieving the goals is in fact harmful *or* side effects of an overwhelming and harmful kind have been overlooked. Classic case: "drug education" programs which aim to and get enrollees

off marijuana and result in getting them on regular cigarettes or alcohol, thereby trading some reduction in (mostly artificial) crimes for far more deaths from lung cancer, cirrhosis of the liver and traffic accidents. (A typical example of ignoring opportunity costs and side effects i.e. bad **GBE**.)

PSYCHOMOTOR SKILLS (Bloom) Learnt muscular skills. The distinction from **cognitive** and **affective** is not always sharp e.g. typing looks psychomotorbut is highly cognitive as well.

PSYCHOLOGICAL EVALUATION PSYCHO-EDUCATIONAL EVALUATION Particular examples of practical evaluation, the first often primarily taxonomical, the second often primarily predictive. The usual standards of validity apply, but are rarely checked; the few studies suggest that even the reliability is very low, and what there is may be largely due to shared bias.

QUALITATIVE (evaluation) A great deal of good evaluation (e.g. of personnel and products) is wholly or chiefly qualitative. But the term is sometimes used to mean "non-experimental" or "not using the quantitative methods of the social sciences," and this has confused the issue, since there is a major tradition and component in evaluation which fits the just-quoted descriptions but is *quantitative*, namely the auditing tradition and the cost analysis component. What has been happening is a gradual convergence of the accountants *and* the qualitative social scientists towards the use of the others' methods *and* the use of some qualitative techniques from humanistic disciplines and low-status social sciences (e.g. ethnography). Obviously evaluation requires all this and more, and the dichotomy between qualitative and quantitative has to be defined clearly and seen in perspective or it is more confusing than enlightening.

QUALITY CONTROL A type of evaluative monitoring, originating in the product manufacturing area, but now used to refer to evaluative monitoring in the human services delivery area. This kind of evaluation is formative in the sense that it is run by the staff responsible for the product, but it is the kind of formative that is essentially "early-

warning summative," because one is endeavoring to ensure that the product, when it reaches the consumer, will appear to be highly satisfactory from the consumer's point of view. Thus quality control is not at all like a common type of evaluative monitoring, which is checking on whether the project is on target; that is a form of **goal-based** evaluation. Quality control should be consumer-oriented evaluation, i.e. **goal-free**, or needs-based evaluation.

QUANTITATIVE (evaluation) Usually refers to the use of numerical analysis methodology from social science or accounting. Cf. **Qualitative.**

QUARTILE (Stat.) See **Percentile.**

QUASI-EXPERIMENTAL DESIGN (Term due to Donald Campbell) When we cannot actually do a *random allocation* of subjects to the control and experimental groups, or cannot arrange that all subjects receive the treatment at the same time, we settle as next best for quasi-experimental design, where we try to simulate a **true experimental** design by *carefully picking* someone or a group for the "control group" (i.e., selecting someone who did not in fact get the treatment but who very closely matches the experimental person/group). Then we study what happens to and perhaps test our "experimental" and "control" groups just as if we had set them up randomly. Of course, the catch is that the reasons (causes) why the experimental group did in fact get the treatment may be because they are different in some way that explains the difference in the outcomes (if there is such a difference), whereas we—not having been able to detect that difference—will think the difference in outcome is due to the difference in the treatment. For example, smokers may, it has been argued, have a higher tendency to lung irritability, an irritation which they find is somewhat reduced in the short run by smoking; and it may be this irritability, not smoking, that yields the higher incidence of lung cancer. Only a "true experiment" could exclude this possibility, but that would probably run into moral problems. However, the weight and web of the quasi-experiments has virtually excluded this possibility. See **Ex Post Facto**.

QUEMAC Acronym for an approach to **metaevaluation**: done by Bob Gowin, a philosopher of education at

Cornell, which emphasizes the identification of unquestioned assumptions in the design. (Questions, Unquestioned Assumptions, Evaluations, Methods, Answers, Concepts.)

QUESTIONNAIRES The basic instrument for surveys and structured interviews. Usually too long, which reduces response rate as well as validity (because it encourages stereotyped or superficial responses.) Must be field-tested; usually a second field-test still uncovers problems e.g. of ambiguity. Interesting problems arise with respect to evaluation questionnaires e.g. what type to use in personnel evaluation when the average response turns out to be a 6 on a 7-point scale, providing inadequate upside discrimination. One can use stronger **anchors**; or rephrase as a **ranking** questionnaire; or impose grading-on-the-curve (Q-sort) methodology, by putting limits on the number of allowable 7's or 6's from any one respondent; or provide deflationary instructions or systems. The first and last of these introduce less distortion where merit levels really are high; the U.S. Air Force once ran into a minor rebellion when it adopted the third alternative. See also **Rating Scales, Symmetry.**

RANDOM A "primitive" or ultimate concept of statistics and probability, i.e., one that cannot be defined in terms of any other except circularly. Texts often define a random sample from a population as one picked in a way that gives every individual in the population an equal probability of being chosen; but one can't define "equal probability" without reference to randomness or a cognate. A distinctly tricky notion. It is not surprising that the first three "tables of random numbers" turned out to have been doctored by their authors; although allegedly generated in (completely different) ways—by mechanical and mathematical procedures—which met the definition just given, they were obviously non-random, e.g. because pages or columns which held a substantial preponderance of a particular digit or a deficit of one particular digit-pair were deleted, whereas of course such pages must occur in any complete listing of all possible combinations. No finite table can be random by the preceding definition. The best definition is relativistic and pragmatic; a choice is random with regard to the variable X if

it is not significantly affected by variables that significantly affect X. Hence a die or cut of cards or turn of the roulette wheel is random with regard to the interests of the players if the number that comes up is caused to do so by variables which are not under the influence of the players' interests.

RANKING, RANK-ORDERING Placing individuals in an order, usually of merit, on the basis of their relative performance on (typically) a test or measurement or observation. Full ranking does not allow ties i.e. two or more individuals with the same rank ("equal third"), partial ranking does; it may then, in the limit case, not be different from **grading.**

RATING Usually same as **grading.**

RATING SCALES Device for standardizing responses to requests for (typically evaluative) judgments. There has been some attempt in the research literature to identify the ideal number of points on a rating scale. An even number counteracts the tendency of some raters to use the midpoint for everything by forcing them to jump one way or the other; on the other hand, it eliminates what is sometimes the only correct response. Scales with 10 or more points generally prove confusing and drop the reliability; with 3 or less (Pass/Not Pass is a two point scale), too much information is thrown away. Five- and (especially) seven-point scales usually work well. It should be noted that the A-F scale is semantically asymmetrical with the usual **anchor** points i.e. it will not give a normal distribution (in the technical sense) of grades for a population in which talent *is* normally distributed.) With + and − and fence-sitting supplements (A+, A,A−,AB,B+,B,B−,BC . . .), it runs to 19 points and with the double + (double −), it has 29 points and becomes essentially ritualistic. Note that the translation of letter grades into numbers, e.g. for purposes of computing a grade-point average, involves assumptions about the equality of the intervals (of merit) between the grades, and about the location of the zero point, which are usually not met (*LE*). See also **Questionnaire.**

RATIONALIZATION Pseudo-justifications, usually provided ex post facto. See **Consonance.**

RATIONALIZATION EVALUATION An evaluation

is sometimes performed in order to provide a rationalization for a predetermined decision. This is much easier than it might appear, and a good many managers know very well how to do it. If they want a program canned, they hire a gunslinger; if they want one salvaged or protected, they hire a sweetheart. Every now and again evaluators are brought in by clients who have got them into the wrong category and the early discussions are likely to be embarassing, annoying or amusing, depending upon how badly you needed the job.

RAW SCORES The actual score on a test, before it is converted into percentiles, grade equivalents, etc.

R&D **R**esearch and **D**evelopment; the basic cyclic (iterative) process of improvement, e.g. of educational materials or consumer products: research, design and prepare, pilot run, investigate (evaluate) results, design improvements, run improved version, etc.

RDD&E **R**esearch, **D**evelopment, **D**iffusion (or **D**issemination) and **E**valuation. A more elaborate acronym for the development process.

REACTIVE EFFECT A phenomenon due to (an **artefact** of) the measurement procedure used: one species of evaluation or investigation artefact. It has two sub-species, content-reaction effects and process-reaction effects. Evaluation-content reactions include cases where a criticism in a preliminary draft of an evaluation is taken to heart by the evaluee and leads to instant improvement, thereby "invalidating" the evaluation. Evaluation-process reactions include cases where the mere occurrence (or even the prospect) of the evaluation materially affects the behavior of the evaluee(s) so that the assessment to be made will not be typical of the program in its pre-evaluated states. Process reactivity is thus content-independent. Although reactive *measurements* have not previously been thus sub-divided, the distinction does apply there and not just to evaluation; but it is less significant. In both cases, unobtrusive approaches may be appropriate to avoid process-reactivity; but on the other hand openness may be required on ethical grounds. The openness may be with respect to content *or* with respect to process *or* both. See **Reasons for evaluation.** Example: **Hawthorne Effect.**

REASONS FOR EVALUATION Two common reasons are to *improve* something (**formative evaluation**) and to make various practical decisions *about* something (**summative evaluation**). Pure interest in determining the merits of something is another kind of summative evaluation. There are also what might be called content-independent reasons for doing evaluation e.g. as a **rationalization** or excuse (for a hatchet job or for funding a favorite) or for **motivation** (to work more carefully or harder). In the excuse case, the *general* nature of the evaluation's content must be known or arranged in advance e.g. by hiring a known "killer" or "sweetheart."

RECOMMENDATIONS In a trivial sense, an evaluation involves an implicit recommendation—that the evaluand be viewed/treated in the way appropriate to the value it was determined to have by the evaluation. But in the specific sense often assumed to be appropriate where "recommendation" is taken to mean "*remedial actions*," evaluations may not lead to them *even if* designed so as to do so (which is much more costly.) That remediation recommendations are not always possible, even when evaluation is possible is obvious in medicine and product evaluation; but because the logic has not been well thought out, it is widely supposed to be a sign of bad design or an absence of humanity when personnel or program evaluations do not lead to them. There are some people who are irremediably incompetent at a given complex task and not even the progress of science will alter that qualitative fact though it may alter percentages. It is a very grave design decision in evaluation to commit a design to producing remedial suggestions, just as it is to undertake to discover **explanations**; it may increase cost and the chance of failure by 1000 percent.

RECOIL EFFECTS When a hunter shoots a deer, he (sic) sometimes bruises his shoulder. Programs affect their staff as well as the clientele. The effect is of secondary importance compared to what happens to the deer or the clientele, but must be included in program evaluation.

REGRESSION TO THE MEAN You may have a run of luck in roulette, but it won't last; your success ration will regress (drop back) to the mean. When a group of subjects is selected for remedial work on the basis of low test scores,

some of them will have scored low only through "bad luck," i.e., the sampling of their skills yielded by (the items on) this test is in fact not typical. If they go through the training and are retested, they will score better simply because any second test would (almost certainly) result in their displaying their skills more impressively. This phenomenon gives an automatic but phony boost to the achievements of "performance contractors" if they are paid on the basis of improvement by the low-scorers. If they had to improve the score of a *random* sample of students, regression *down* to the mean would offset the regression *up* to the mean we have just discussed. But they are normally called in to help the students who "need it most" and picking that group by testing will result in including a number who do *not* need help. (It will also *exclude* some who *do.*) *Multiple or longer tests or the addition of teacher (expert judge) evaluations reduce this source of error.*

RELATIVISM/SUBJECTIVISM Roughly speaking, the view that there is no objective reality about which the evaluator is to ascertain the truth, but only various perspectives or approaches or responses, amongst which selection is fairly arbitrary or is dependent upon aesthetic and psychological considerations rather than scientific ones. The contrary point of view would naturally be referred to as absolutism or objectivism; in one technical sense used in philosophy the opposite of subjectivism is called the doctrine of realism. The fundamental logical fallacy that confounds many discussions of this issue is the failure to see the full implications of the fact that relativism is a self-refuting doctrine, i.e., "relativism is true" can be no more true than "relativism is false," and hence relativism can hardly represent a Great Truth, since it is self-refuting. One very important implication of this point for evaluation practice is as follows: in a situation where a number of different approaches, methodologies or perspectives on a particular program (for example) are possible and all are about equally plausible, it does *not* follow that any one of them would constitute a defensible evaluation. The only thing that follows is that giving *all* of them *and* the statement that all of them are equally defensible, would constitute a defensible evaluation. The moment that one has seen that alternative approaches are equally good, although they yield incom-

patible results, one has seen that no one of these can be thought of as sound in itself, *just because* the assertion of any one of them implies the denial of the others and that denial is, in such a case, illegitimate. Hence the assertion of any one of them by itself is illegitimate. If, on the other hand, the different positions are *not* incompatible, then they must still be given in order to present a comprehensive picture of whatever is being evaluated. In neither case, then, is giving a single one of these perspectives defensible. In short, the great difficulties of establishing one evaluation conclusion by comparison with others cannot be avoided by arbitrarily picking one, but only by proving the superiority of one or including all *as* perspectives, a term which correctly implies the existence of a reality which is only partly revealed in each view. Thus it converts incompatible reports into complementary ones i.e. it converts relativism into objectivism. Merely giving several apparently incompatible accounts in an evaluation is incompetent; showing how they can be reconciled i.e. seen *as* perspectives is also required. (Or else a proof that there is no single reality.) The presupposition that there is a single reality is not an arbitrary one, any more than the assumption that the future will be somewhat like the past is arbitrary; these are well-established. Determinism was equally well-established and we have now had to qualify it slightly because of the Uncertainty Principle. We have not yet encountered good reasons for qualifying the assumptions of realism and induction (the technical names for the two previously mentioned.)

As the practical end of these considerations, it must be recognized that even evaluations ultimately based on "mere preferences" may still be completely objective. One must distinguish sharply between the fact that the ultimate basis of merit in such cases is mere preference, on which the subject *is* the ultimate source of authority, and the fallacy of supposing that the subject must therefore be the ultimate source of authority about the *merits* of whatever is being evaluated. Even in the domain of pure taste, the subject may simply not have researched the range of options properly, or avoided the biassing effects of labels and advertising, orrecommendations by friends, so the evaluator may be able to identify critical competitors that outperform the subject's favorite candidate, *in terms of the subject's own taste.* And of course identifying Best Buys for an individual in-

volves a second dimension (cost) which the evaluator is often able to determine and combine more reliably than the amateur. The moment we move the least step from areas where superiority is unidimensional, instantaneous, and entirely taste-dependent, then we find the subject beginning to make errors of synthesis in putting together two or three dimensions of preference (halo or sequencing effects, for example), or in extrapolating to continued liking, errors that an evaluator can reduce or eliminate by appropriate experimental design, often leading to a conclusion quite different from that which the subject had formed. One step further away, and we find the possibility of the subject making first-level errors of judgment, e.g. about what they need (or even what they want) by contrast with what they like, and these can certainly be reduced or eliminated by appropriate evaluation design. In the general case of the evaluation of consumer goods, the question of whether one can identify "the best" product with complete objectivity, despite a substantial range of different interests and preferences at the basic level by the relevant consumer group, is simply a question of whether the interproduct variations in performance outweigh the interconsumer variations in preference. Enormous variations in preference may be completely blotted out by the tremendous superiority of a single product over another, such that it "scores" so much on several dimensions which are accorded significant value by all the relevant consumers, that even the outlandish tastes (weightings) of some of the consumers with respect to some of the other dimensions cannot elevate any of the competitive products to the same level of total score, even for those with the atypical tastes. Thus huge interpersonal differences in *all the relevant* preferences do *not* demonstrate the relativism of *evaluations* which depend on them.

RELIABILITY (Stat.) Reliability in the *technical* sense is the *consistency* with which an instrument or person measures whatever it is designed to assess. If a thermometer always says 90 degrees Centigrade when placed in boiling distilled water at sea-level, it is 100% "reliable," though inaccurate. It is useful to distinguish test-retest reliability (the example just given) from interjudge reliability (which would be exhibited if several thermometers gave the same reading). There are many psychological tests which are

test-retest reliable but not interjudge (i.e., inter-administrator) reliable: the reverse is less common. In the *everyday* sense, reliability means the same as the technical term **validity**; we'd say that a thermometer which reads 90 degrees Centigrade when it should read 100 degrees Centigrade wasn't very reliable. This confusing situation could easily have been avoided by using the term "consistency" instead of introducing a technical use of "reliability" but that was in the days when jargon was thought to be a sign of scientific sophistication. As it is, reliability is a necessary but not a sufficient condition for validity, hence worth checking first since in its absence validity can't be there. (There is, unfortunately, a hyper-technical exception to this.)

RELIABILITY (of evaluation) A largely unknown quantity, easily obtained by running replications of evaluations; either serially or in parallel. The few data on these make clear that reliability (apart from spurious effects such as shared bias) is not high. The use of **calibration** exercises and **checklists** and trained evaluators can improve this enormously.

REMEDIATION A specific recommendation for improvement, characteristic of—and certainly desirable in—formative rather than summative evaluation. But **formative** can be useful without any remediation suggestions, and it is in general more difficult (sometimes completely impossible) and more expensive if it aims for remediation. See also **Recommenation.**

REPLICATION A very rare phenomenon, contrary to reports, mainly because people do not take the notion of serious testing for implementation (e.g. through the use of an **index of implementation**) as an automatic requirement on any supposed replication. Even the methodology for replication is poorly thought out; for example, whether the replicator should have any detailed knowledge of the results of the primary site? Such knowledge is seriously biasing—on the other hand, it significantly simplifies the preparations for ranges of measurement, etc. It is probably quite important to arrange at least some replications where the (e.g.) program to be replicated is simply described in operational terms, perhaps with the incidental remark that it has shown "promising results" at the primary site.

REPORT WRITING/GIVING One of several areas in evaluation where creativity and originality are really important, as well as knowledge about **diffusion** and **dissemination**. Reports must be tailored to **audience** as well as client needs and may require a minor **needs assessment** of their own. Multiple versions, sometimes using different media, as well as different vocabularies, are often appropriate. Reports are products and should be looked at in terms of the **KEC**—field-testing them is by no means inappropriate. Who has time and resources for all this? It depends whether you are really interested in **implementation** of the evaluation. Would you write it in Greek? No, so why *assume* that you are not writing it in the *equivalent* of Greek, as far as your audiences are concerned?

RESEARCH The general field of disciplined investigation, covering the humanities, the sciences, jurisprudence, etc. Evaluation research is one subdivision; there is no way to distinguish other research from evaluation (apart from content) except by distorting one or the other. "Evaluation research" is usually just a self-important name for serious evaluation; it would be better used to refer to research *on* evaluation methodology, or research that pushes out the frontiers of evaluation, or at least research that involves considerable investigatory difficulty or originality. Cf. performing arts vs. creative arts.

RESEARCH INTEGRATION, RESEARCH SYNTHESIS See **Meta-analysis.**

RESEARCH EVALUATION Evaluating the quality and/or value and/or amount of research (proposed or performed) is crucial for e.g. funding decisions and university personnel evaluation. It involves the **worth/merit** distinction—"worth" here refers to the social or intellectual payoffs from the research, "merit" to its intrinsic (professional) quality. While some judgment is always involved, that is no excuse for allowing the usually *wholly* judgmental process; one can quantify and in other ways objectify the merit and worth of almost all research performances *to the degree requisite* for personnel evaluation.

RESPONSIVE EVALUATION Bob Stake's current approach, which contrasts with what he calls "preordinate" evaluation, where there is a predetermined evaluation de-

sign. In responsive evaluation, one picks up whatever turns up and deals with it as seems appropriate, in the light of the known and unfolding interests of the various audiences. The emphasis is on rich description, not testing. The risk is of course a lack of structure or of valid proof, but the trade-off is the avoidance of the risk of a preordinate evaluation—a rigid and narrow outcome of little interest to the audiences. Cf. **Evaluation-Specific Methodology, Naturalistic Evaluation**.

RESPONSE SET Tendency to respond in a particular way, regardless of the merits of the particular case. Some respondents tend to rate everything very high on a scale of merit, others rate everything low, and yet others put everything in the middle. One can't argue out of context that such patterns are incorrect; there are plenty of situations in which those are exactly the correct responses. When we're talking about response set, however, we mean the cases where these rigid response patterns emerge from general habits and not from well thought-out consistency.

RESPONSIBILITY EVALUATION Evaluation that is oriented e identification of the responsible person(s) or the degree of responsibility, and hence usually the degree of culpability or merit. Responsibility has causality as a necessary but not a sufficient condition. Culpability similarly presupposes responsibility but involves further conditions from ethics. Social scientists like most people not trained in the law or casuistry are typically totally confused about such issues e.g. supposing that evaluations shouldn't be done (or published) because "they may be abused." The abuse is culpable; but it is *failure* to publish (assuming it's professional-quality work of some *prima facie* intellectual or social value) that would be culpable (e.g. the Jensen case). A different kind of example involves keeping really bad teachers on in a school district *because* the alternative of attempting dismissal involves effort, is unpopular with the union, and usually unsuccessful. The responsibility is to the pupils who are sacrificed at the rate of 30 per annum per bad teacher; and that responsibility is so serious that you (the superintendent or the board) have to try for removal because (a) you *may* succeed, (b) the effects may be *on balance* good, (c) you may learn how to do it better next time. The evaluation of schools should (normally) only be done in

terms of the variables over which the school has control—in the short run and often in the long run, this does *not* include scores on standardized tests. (See **SEP**). The evaluation of evaluations should never be done in terms of results, because the evaluator is not responsible for implementation; but it *should* be done in terms of results *if* implemented. Ref. ***Primary Philosophy***, Scriven, McGraw-Hill, 1966.

RETURN ON INVESTMENT (ROI) One of the measures of merit or worth in fiscal evaluation; usually quoted as a per annum percentage rate.

RISK(S), EVALUATING The classic expectancy approach in which the products of the probability of each outcome by its utility are compared, thus converting the two dimensions of risk and utility into the one (of expectancy), has certain weaknesses. For example, it ignores the variable value of risk itself to different individuals; the gambler likes it, many others seek to minimize it. "Risk management" is a topic that has begun to appear with increasing frequency in planning and management training curricula. One reason that evaluations are not implemented is because the evaluator has failed to see that risks have different significance for implementers by contrast with consumers; a program or policy (etc.) which should be implemented, in terms of its probable benefit to the consumers may be one which carries a high risk for the implementers, because their reward schedule is often radically different from that of the consumer (usually as a result of bad planning and management at a higher level. Two classic examples are the classification of documents as Top Secret and the hiring of personnel about whom there is a breath of suspicion; in each situation, the implementer gets zapped by review panels exercising 100 percent hindsight after a disaster if there is the least trace of a negative indicator, and in neither case is there ever a reward for taking a reasonable risk—in fact, there's never a review panel. Consequently, the public's utilities are *not* optimized and are often reversed. The present political-plus-media environment in the U.S. may be one in which the risk configuration for the road to the Presidency (or the legislature) is so different from that required to do the job right as to guarantee the election of poor incumbents who were great candidates.

RIGHT-TO-KNOW The legal domain of impacted populations' access to information; much increased lately e.g. through "open file" legislation.

RHETORIC, THE NEW The title of a book by C. Perelman and L. Olbrechts-Tyteca (Notre Dame, 1969), which attempted to develop a new logic of persuasion, reviving the spirit of pre-Ramist efforts. (Since Ramus (1572), the view of rhetoric as the art of empty and illogical persuasion has been dominant; the concept of "logical analysis," as separate from rhetoric is Ramist.) This area is of the greatest importance to evaluation methodology as Ernest House has stressed (e.g. in ***Evaluating with Validity***, Sage, 1980), because of the extent to which evaluations have—whether intentionally or not—the function of persuasion and not just reporting. The New Rhetoric emerged from the context of studying legal reasoning where the same situation obtains and was poorly recognized. The same push for reappraisal and new models has occurred in logic (see ***Informal Logic,*** eds. Blair and Johnson, Edgepress, 1980), and in the social sciences with the move towards naturalistic methodology. It is all part of the backlash against neo-positivist philosophy of science and the worship of the Newtonian model of science. Evaluation's fate clearly lies with the new movements.

RITUAL(ISTIC) EVALUATION One of the reasons for doing evaluation that has nothing to do with the content of the evaluation (and hence is unlike **formative** and **summative**—or **rationalization**—evaluation) is the ritual function i.e. the doing of an evaluation because it is required, although nobody has the faintest intention of either doing it well or taking any account of what it says. Evaluators are quite often called in to situations like this, although they may not even be recognized as cases of ritual evaluation by the client. (Evaluation in the bilingual education area is currently mostly ritualistic.) It is an important part of the preliminary discussions in serious evaluation to get clear exactly what kind of implementation is planned, under various hypotheses about what the content of the evaluation report might be; unless, of course, you have time to spare, need the money, and are not misleading any remote audiences. The third condition essentially never applies. See also **Motivational Evaluation.**

ROBUSTNESS (Stat.) Statistical tests and techniques depend to varying degrees on assumptions especially about the population of origin. The *less* they depend on such assumptions, the *more* robust they are. The t-test assumes normality, non-parametric ("distribution-free") statistics are often considerably more robust. One might translate "robust" as "stable under variation of conditions." The concept is also applicable to and most important in the evaluation of experimental designs and meta-evaluation. Designs should be set up to give *definite* answers to at least some of the *most important* questions no matter how the data turns out, a matter quite different from their cost-effectiveness, power, or elegance (the latter is a kind of limit case of efficiency or power.) Evaluations should be set up so as to "go for the jugular" i.e. get an adequately reliable answer to the key evaluative question(s) *first*, adding the trimmings later *if* nothing goes wrong with Part One. This affects budget, staff and time-line planning. And it has a cost as does robustness in statistics; for example, robust approaches will not be maximally elegant if everything goes right. But meta-evaluation will normally show that a minimax approach is called for, which means robust evaluation.

ROLE (of evaluator) The evaluator plays more roles than Olivier, or should. Major ones include therapist/confessor, educator, arbitrator, co-author, "the enemy," trouble-shooter, jury, judge, attorney.

RORSCHACH EFFECT An extremely complex evaluation, if not carefully and rationally synthesized into an executive summary report, provides a confusing mass of positive and negative comments, and the unskilled and/or strongly biased client can easily project onto ("see in," rationalize from) such a backdrop whatever perceptions s/he originally had.

RUBRIC Scoring *or* grading *or* (conceivably) ranking key for a test.

SALIENCE SCORING The practice of requesting respondents to use only those scales which, they felt, most significantly influenced them. It focuses attention on the most important features of whatever is being rated, and it

greatly reduces processing time.

SCALES See **Measurement**.

SCOPE OF WORK This is the part of an RFP or a proposal which describes exactly what is to be done, at the level of description which refers to the activities as they might be seen by a visitor without special methodological skills or insight, rather than to their goals, achievements, process or purpose. In point of fact, scope of work statements tend to drift off into descriptions that are somewhat less than observationally testable. The scope of work statement is an important part of making accountability possible on a contract, and is therefore an important part of the specifications in an RFP or a proposal.

SCORING Assigning numbers to an evaluand, (usually a performance) usually from an interval scale i.e. one in which the points all have equal value. Sometimes numbers are used as grades without commitment to **point constancy**, but this is misleading—letters should be used instead, and the attempt to convert them to numbers e.g. to calculate GPAs should be protested unless point constancy holds at least to an approximation that will not yield errors (*LE*.) Usually tests should be impressionastically graded as well as scored, both to get the cutting scores and to provide insurance against deviations from point constancy. Scoring not only requires point constancy but also serious consideration of the definition of a zero score: no answer? hopelessly bad answer? both? ("both" is a hopelessly bad answer.)

SECONDARY ANALYSIS Reassessment of an experiment or investigation, either by reanalysis of the data or reconsideration of the interpretation. Gathering new data would normally constitute **replication**; but there are intermediate cases. Sometimes used to refer to reviews of large numbers of studies; See **Meta-analysis, Secondary Evaluation.**

SECONDARY EVALUATION (Cook) Reanalysis of original—or original plus new—data in order to produce a new evaluation of a particular project (etc.). Russell Sage Foundation commissioned a series of books in which famous evaluations were treated in this way, beginning with

different and more accurate response to the next (or any later) question.

SES Socio-Economic Status.

SENSORY EVALUATION Wine-tasting when done scientifically, the better restaurant reviews, the Consumers Union report on bottled water, remind us of the important difference between dismissing something as a "mere matter of taste" and doing sensory evaluation which does not *eliminate* dependence on performance but improves its reliability and improves the evaluative inference e.g. by eliminating distractors (such as labels), using multiple independent raters and standardized sets of criteria.

SHARED BIAS The principal problem with using multiple expert opinion for validation of evaluations is that the agreement (if any) may be due to common error; obvious and serious examples occur in peer review of research proposals, where the panelists tend to reflect current fads in the field to the detriment of innovators, and in **accreditation.** The best antidote is often the use of intellectually and not just institutionally external judges e.g. radical critics of the field. "The inference from reliability to validity must bridge the chasm of shared bias."

SIDE EFFECTS Side effects are the unintended good and bad effects of the program or product being evaluated. Sometimes the term refers to effects that were intended but are not part of the goals of the program e.g. employment of staff. In either case, they may or may not have been expected, predicted or anticipated (a minor point). In the **Key Evaluation Checklist** a distinction is made between side effects and standard effects on impacted non-target populations, i.e. side-*populations*, but both are often called side effects.

SIGNIFICANT, SIGNIFICANCE The overall, synthesized conclusion of an evaluation; may relate to social or professional or intellectual significance. Statistical significance, when relevant at all, is one of a dozen necessary conditions for real significance. The significance of an intervention may be considerable even if it had no effects in the intended directions which might be cognitive or health gains; it may have employed many people, raised general

Tom Cook's secondary evaluation of Sesame Street. Extremely important because (a) it gives potential clients some basis for estimating the reliability of evaluators (in the case just cited, the estimate would be fairly low); (b) it gives evaluators the chance to identify and learn from their mistakes. Evaluations have all too often been fugitive documents and hence have not received the benefit of later discussion in "the literature" as would a research report published in a journal; a weakness in the field. (Similar problem applies to classified material). Cf. **Metaevaluation.**

SECRET CONTRACT BIAS In proposal and personnel review, raters are often too lenient because they know that the roles will be reversed on another occasion and they think or intuit that if everyone sees that, and acts accordingly, "we'll all come out smelling like roses." Typical unprofessional conduct typical of the professions. A good counterbalance is to rate everyone on the long-term validity of their ratings.

SEMI-INTERQUARTILE RANGE (Stat.) Half the interval between the score that marks the top score of the lowest or first quartile (i.e. the lowest quarter of the group being studied, after they have been ranked according to the variable of interest, e.g. test scores), and the score that marks the top of the third quartile. This is a useful measure of the range of a variable in a population, especially when it is not a normal distribution (where the "standard deviation" would be used). It amounts to averaging the intervals between the median and the individuals who are halfway out to the ends of the distribution, one in each direction. Thus it is not affected by oddities occurring at the extreme ends of the distribution.

SEP (School Evaluation Profile) An instrument for evaluating the performance of schools (and hence districts, principals etc.), which looks only at variables the school (population) controls. Available (in field test form only) from Edgepress. See **Responsibility Evaluation.**

SEQUENCING EFFECT The influence of the order of items (tests etc.) upon responses; it jeopardizes the test's validity when items are removed e.g. for racial bias, since the item might have preconditioned the respondent (in a way that has nothing to do with its bias) so as to give a

awareness of problems, produced other gains. The absence of *overall* significant effects may also be due to dilution of good effects in a pool of poor programs producing no effects: one cannot infer from an overall-null to individual-nulls. For this reason, "lumping designs" are much less desirable than "splitting designs" in which separate studies are made of many sites or sub-treatments (see **Replication, Meta-analysis.)** Omega-statistics and Glass' "standardized effect size" are attempts to produce measures that more nearly reflect true significance than does the p level of the absolute size of the results.

SIMULATIONS Re-creations of typical job situations to provide a realistic test of aptitudes or abilities. See **Clinical Performance Testing.**

SMILES TEST (of a program) People like it.

SOCIAL INDICATOR See **Indicator.**

SOCIAL SCIENCE MODEL (of evaluation) The (naive) view that evaluation is an application of standard social science methodology. See **Evaluation**.

SOFT (approach to evaluation). Uses **implementation** data or the **Smiles Test.** See **Hard.**

SOLE SOURCE "Sole-sourcing" a contract is an alternative to "putting it out to bid," via publishing an RFP. Sole sourcing is open to the abuse of the contract officer from the agency letting contracts to his or her buddies without regard to whether the price is excessive or the quality unsatisfactory; on the other hand, it is very much faster, it costs less if you take account of the time for preparing RFP's and proposals in cases where a very large number of these would be written for a very complex RFP, and it is sometimes mandatory when it is provable that the skills and/or resources required are available from only one contractor within the necessary time-frame. Simple controls can prevent the kind of abuse mentioned.

SPEEDED (tests) Also called power tests, those tests with a time *limit* (the time taken by each individual is usually *not* recorded, though it is in *timed* tests). These are often better instruments for evaluation or prediction than the same test would be with no time limit—usually because the

criterion behavior involves doing something under time pressure, but sometimes, as in IQ tests, just as a matter of empirical fact. A test is sometimes defined as speeded if only 75 percent of the testees finish in time.

SPONSOR (of evaluation) Whoever or whatever funds or arranges it: referred to as "instigator" in KEC.

STAKEHOLDER An interested party in an evaluation e.g. a politician who supported the original program.

STANDARD(S) The performance associated with a particular rating or "grade" on a given criterion or dimension of achievements; e.g. 80 percent success may be the standard for passing the written portion (dimension) of the driver's license test. A cutting score defines a standard, but standards can be given in non-quantitative grading contexts, e.g. by providing exemplars, as in holistic grading of composition samples.

STANDARD DEVIATION (Stat.) A technical measure of dispersion; in a normal distribution, about two thirds of the population lies within one standard deviation of the mean, median, or mode (which are the same in this case.) The S.D. is simply the mean of the sum of the squares of the deviations i.e. distances from the mean.

STANDARD ERROR OF MEASUREMENT (Stat.) There are several alternative definitions of this term, all of which attempt to give a precise meaning to the notion of the intrinsic inaccuracy of an instrument, typically a test.

STANDARD SCORE Originally, scores defined as deviations from the mean, divided by the standard deviation. (**Effect Size** is an example.) More casually, various linear transformations of the above (Z-scores) aimed to avoid negative scores.

STANDARDIZED TEST Standardized tests are ones with standardized instructions for administration, use, scoring and interpretation, standard printed forms and content, and often with standardized statistical properties, that have been validated on a large sample of a defined population. They are *usually* norm-referenced, at the moment, but the terms are not synonymous since a criterion-referenced test can also be standardized. Having the norms (etc.) on a

test does mean it's standardized in one respect, but it does not mean it's *just* a norm-referenced test in the technical sense; it may (also) be criterion-referenced, which implies a different technical approach to its construction and not just a different purpose.

STANINES (or stanine scores). If you are perverse enough to divide a distribution into nine equal parts instead of ten (see **decile**), they are called stanines and the cutting scores that demarcate them are called stanine scores. They are numbered from the bottom up. See also **Percentiles.**

STATISTICAL SIGNIFICANCE (Stat.) When the difference between two results is determined to be "statistically significant," the evaluator can conclude that the difference is probably not due to chance. The "level of significance" determines the degree of certainty or confidence with which we can rule out chance (i.e. rule out the "null hypothesis"). Unfortunately, if very large samples are used even tiny differences become statistically significant though they may have no educational value at all. Omega statistics provide a partial correction for this. Cf. **Interocular Difference.**

STEM The text of a multiple-choice test item that precedes the listing of the possible responses.

STRATIFICATION A sample is said to be stratified if it has been deliberately chosen so as to include an appropriate number of entities from each of several population subgroups. For example, one usually stratifies the sample of students in K-12 educational evaluations with regard to gender, aiming at 50 percent males and 50 percent females. If one selects a random sample of females to make up half of the experimental and half of the control group and a random sample of males for the other half, then one has a "stratified random sample." If you stratify on too many variables you may not be able to make a random choice of subjects in a particular stratum—there may be no or only one eligible candidate. If one stratifies on very few or no variables, one has to use larger random samples to compensate. Stratification is only justified with regard to variables that probably interact with the treatment variable, and it only increases efficiency, not validity, unless you do it in addition to using large numbers i.e. abandon the efficiency

gains it makes possible. Indeed it runs some risk of reducing validity because you may *not* cover a key variable (through ignorance) and your reduced sample size may not take care of it.

STRENGTHS ASSESSMENT Looking at resources available, including time: it defines the range of the possible and hence is important in both needs assessment and the identification of critical competitors, as well as in making remediation suggestions.

STYLE RESEARCH Investigations of two kinds; *either* descriptive investigations of the actual stylistic characteristics of people in e.g. certain professions such as teaching or managing; *or* investigations of the correlations between certain style characteristics and successful outcomes. The second kind of investigation is of great importance to evaluation, since discoveries of substantial correlations would allow certain types of evaluation to be performed on a process basis, which currently can only be done legitimately by looking at outcomes. (However, **personnel evaluation** could *not* be done in that way, even if the correlations were discovered.) The former kind of investigation—a typical example is studying the frequency with which teachers utter questions by comparison with declarative sentences or commands—is pure research, and extremely hard to justify as of either intellectual or social interest unless the second kind of connection can be made. In general, style research has come up with disappointingly few winners. (Actual Learning Time is probably the most important and possibly the only exception.) No doubt the interactions between the personality, the style, the age and type of recipient and the subject matter prevent any simple results; but the poor results of research on interactions suggests that the interactions are so strong as to obliterate even very limited recommendations. We must instead fall back to treating positive results as *possible remedies, not probability indicators of merit.*

SUMMATIVE EVALUATION Summative evaluation of a program (etc.) is conducted *after* completion and *for* the benefit of some *external* audience or decision-maker (e.g. funding agency, or future possible users,) though it may be *done* by either internal or external evaluators or a mixture. For reasons of credibility, it is much more likely to involve

external evaluators than is a formative evaluation. Should not be confused with **outcome evaluation**, which is simply an evaluation *focused on* outcomes rather than on process—it could be either formative or summative. (This confusion occurs in the introduction to the ERS ***Evaluation Standards***, 1980 Field Edition). Should not be confused with *holistic* evaluation—it may be holistic or *analytic*.

SUPERCOGNITIVE The domain of performance on cognitive (or information/communication) skills that is a quantum jump beyond normal levels, e.g. speed reading, lightning calculating, memory mastery, speedspeak or fastalk, tri-linguality, shorthand. Cf. **Hypercognitive**

SURVEY METHODS (in evaluation) See **Evaluation-Specific Methodology**.

SYMETTRY of evaluative indicators. It is a common error to suppose (or unwittingly to arrange) that the converse or absence of an indicator of merit is an indicator of demerit. This is illustrated by the assumption that items in evaluative questionnaires can be rewritten positively or negatively to suit the configural requirement of foiling stereotyped responses. But "Frequently lies" is a strong indicator of demerit, while "Does not frequently lie" is not even a weak indicator of (salient) merit. (Salient merit i.e. commendable behavior is what one rewards, not "being better than the worst one could possibly be.") The preceding is an epistemological point about symmetry (related to the virtue/supererogation distinction in ethics). There are also methodological asymmetries; for example, an item requesting a report on absences e.g. "Was sometimes absent without leave" can be answered affirmatively by respondents who were often not there themselves but who observed one or more such absences by the evaluee; but "Was rarely absent without leave" will be checked "Don't know" by the same respondents since it calls for knowledge they do not possess.

SYNTHESIS (of studies) The integration of multiple research studies into an overall picture is a field which has recently received considerable attention. These "reviews of the literature" are not only evaluations in themselves, with—it turns out—some quite complex methodology and viable alternatives involved on the way to a bottom line; but

they are also a key element in the evaluator's repertoire since they provide the basis for identifying e.g. critical competitors and possible side-effects. See **Meta-analysis.**

SYNTHESIS (in evaluation) The process of combining a set of ratings on several dimensions into an overall evaluation. Usually necessary and defensible; sometimes inappropriate because it requires a decision on relative weighting which *sometimes* is impossible. Those occasions require giving just the ratings on the separate dimensions. It is desirable to require an explicit statement and justification of the synthesis procedure since this will often expose: (a) arbitrary assumptions, (b) inconsistent applications. In the evaluation of faculty, for example, the *de facto* weighting of research vs. teaching is often nearer to 5:1 in institutions whose rhetoric claims parity; but it may vary widely between departments or between successive chairs in the same department. The evaluation of student course work by the letter grade is often cited as an example of indefensible synthesis; in fact it is a perfectly defensible summative evaluation, though it is unjustifiable for formative feedback to the student. "Synthesis by salience summary" illustrates another trap; a teacher is rated on 35 scales by students and the printout only shows cases of statistically significant departures of the ratings from the norms. This seems plausible enough; but since the dimensions have not been independently validated, (and are not independent) it not only involves focusing on style characteristics which are being appraised on *a priori* grounds, but it also involves all the confusions of **ranking** instead of grading. The importance of correct synthesis is illustrated by a psychiatrist on the staff at the University of Minnesota who became legendary for requesting a grant so that a graduate student could "pull his research results together"; his "research results" being a complete set of taped recordings of five years of therapy. Evaluators that are tempted to "turn the facts over to the decision-makers, and let *them* make the value-judgments" should remember that evaluations are interpretations that require *all* the professional skills in the repertoire; a scientist's role does not end with observation and measurement. Weighted-sum synthesis is linear synthesis and usually works well. Rarely, as in the evaluation of backgammon board positions or in evaluating patients on the MMPI, we

need non-linear synthesis rules. Synthesis is perhaps the key cognitive skill in evaluation; it covers evaluating invoked by the phrase "balanced judgment" as well as the **apples and oranges** difficulties. Its cousins appear in the core of all intellectual activity; in science, not only in theorizing and identifying the presence of a theoretical construct from the data but in **research synthesis.** In evaluation, the wish to avoid it manifests itself in laissez-faire evaluation's extreme forms of the **naturalistic approach**. Balking at the final synthesis is often (not always) balking at the value judgment itself and close to **valuephobia.**

SYSTEMS ANALYSIS The term is generally used interchangeably with "systems approach" and "systems theory." This approach places the product or program being evaluated into the context of some total system. Systems analysis includes an investigation of how the components of the program/product being evaluated interact and how the environment (system) in which the program/product exists affect it. The "total system" is not clearly defined, varying from a particular institution to the universe at large, hence the approach tends to be more an orientation than an exact formula and the results of its use range from the abysmally trivial to deep insights. (Ref. C.W. Churchman, ***The Design of Inquiring Systems***).

TA Technology Assessment. An evaluation, particularly with respect to probable impact, of usually new) technologies. Discussed in more detail under **Technology** below.

TARGET POPULATION The intended recipients or consumers. Cf. **Impacted Population**.

TAXONOMIES Classifications, most notably Bloom's taxonomy of educational objectives; a huge literature has grown up around these taxonomies, which are rather simplistic in their assumptions and excessively complex in their ramifications.

TEACHING TO THE TEST The practice of teaching just or mostly those skills or facts that will be tested, based on illicit prior knowledge of or inference as to the test content. If the test is fully comprehensive, e.g. testing

knowledge of the "times tables" by calling for all of them, this is simply task-orientation and no crime. But most tests only sample a domain of behavior and generalize from performance on that sample as to overall performance in the domain, and *that generalization* will be erroneous when teaching to the test has occurred. A serious weakness of teacher-constructed tests is that they create the same situation *ex post facto*: see **Testing to the Teaching.**

TECHNOLOGY ASSESSMENT A burgeoning form of evaluation which aims to assess the total impact of (typically) a new technology. A cross between futurism and systems analysis and consequently done at every level from ludicrously superficial to brilliant. OTA usually scores well above the middle of the possible range. The process remains in need of systematization; predicting that cassette recorders would displace books was clearly fallacious at the time, while predicting that hand-held optical-scanning voice-input/output printing micro-computers will virtually eliminate the necessity for instruction in basic skills by 1990 seems now (1980) to be so certain that the vast restructuring of the educational system which it entails should have long begun. One good feature of TA futurism might seem to be that in the long run we'll know who was right; but so much of it relates to *potential* that refutation is hard.

TERROR The effect frequently induced by goal-free evaluation (sometimes by the *thought* of it) in the whole cast of actors—evaluators, managers, evaluees. The "terror test" is the use of this awful threat to determine whether the cast is competent.

TESTS (& TEST ANXIETY). Tests are poor instruments when the subjects are *more* anxious than they would be in the criterion situation or when they test a domain poorly matched to the test's alleged domain; but they are better than most observers including the classroom teacher in many, many cases.

TESTING TO THE TEACHING Designing tests to measure just what is actually taught instead of testing learning in the domain about which conclusions will be or need to be drawn. Tests of a reading program that only use words actually covered in class will give a false picture of reading skills. As with "teaching to the test," this situation will not

be improper in the extreme case where the teaching covers the whole domain.

TEST WISE Said of a subject who has acquired substantial skills in test-taking e.g. learning to say False on all items which say "always" or "never," or (to give a sophisticated example) learning not to check answers on items one hasn't time to read carefully, *if* a "**correction for guessing**" is being used, but to do so if it is not.

THEORIES A luxury for the evaluator, since they are not even essential for explanations, and explanations are not essential for (99 percent of all) evaluations. It is a gross though frequent blunder to suppose that "one needs a theory of learning in order to evaluate teaching". One does not need to know anything at all about electronics in order to evaluate electronic typewriters, even formatively, and having such knowledge often *adversely* affects a summative evaluation. See **Conceptual Scheme.**

THERAPEUTIC ROLE (OR MODEL) OF THE EVALUATOR The very nature of the evaluation situation creates pressures that sometimes mold it into a therapist-patient or group therapy interaction; this is particularly but not only true with regard to external evaluation. First, there is—in such a case—the client's feeling of having exhausted his/her own resources, needing help badly, perhaps desperately. Second, there is the aura of expertise and esoterica which (sometimes) surrounds the external expert. Third, there are the technical diagnoses and magical rites prescribed by the good doctor. Since it's doubtful that there is in general much more to psychotherapy than this, an amalgam which is enough to generate at least the placebo effect, the analogy is clear—and should be disturbing. The main problem with placebo and Hawthorne effects is their transitory nature and the evaluator who fades back into the hills after an ecstatic client's testimonial dinner may have to sneak back for a look around a year later if s/he wants to get a good idea of whether the recommendations were (a) solutions to the problems, (b) adopted, (c) supported. Hence follow-up studies, sadly lacking in psychotherapy research (or innovation evaluation) and often devastating when done, are just as important in meta-evaluation.

TIME DISCOUNTING A term from fiscal evaluation

which refers to the systematic process of discounting future benefits, e.g. income, for the fact that they *are* in the future and hence (regardless of the *risk*, an essentially independent source of value reduction for merely probable future benefits) lose the earnings that those monies would yield if in hand now, in the interval before they will in fact materialize. Time discounting can be done with reference to any past or future moment as base point, but is usually done by calculating everything in terms of "true present value."

TIME MANAGEMENT An aspect of management consulting with which the "general practitioner" evaluator should be familiar; it ranges from the trivial to the highly valuable. Psychologists from William James to B.F. Skinner are amongst those who have made valuable contributions to it and it *can* yield very substantial output gains at very small cost both for the evaluator and for clients or evaluees. It was James who suggested listing tasks to be done in decreasing order of enjoyability and beginning at the *bottom*, perhaps since that gives you the largest reduction of guilt and the biggest gain in charm for the remaining list. (Ref. McCay, James ***The Management of Time***, Prentice Hall.)

TIME SERIES (See **Interrupted Time Series Analysis)**

TRAINING OF EVALUATORS Evaluators, like philosophers, and unlike virtually every other kind of professional, should be regarded as having a general obligation to know as much as possible about as much as possible. While it is feasible and indeed quite common for evaluators to specialize either in particular methodologies or in particular subject matter areas, the costs of doing this are usually rather obvious in their work. It is probably a consequence of the relative youth of evaluation as a discipline that the search for illuminating analogies from other disciplines is still so productive; but the other reason for versatility will always be with us, namely that it enables one to do better as an evaluator in as wide a range of subject matter areas as possible. Columbia University used to have a requirement that students could not be accepted for the doctorate in philosophy unless they had a Master's degree in another subject, and an analogous requirement might be quite desirable in evaluation. However, it is commonly asserted that the preliminary degree should be in statistics, tests, and

measurement. The problem with that requirement is that it leads to a strong methodological bias in the eventual practice of the professional. While skill in the quantitative methodologies is highly desirable, it does not have to be a *preliminary* to evaluation training; the reverse sequence may be preferable. A simple formula for becoming a good evaluator is to learn how to do everything that is required by the Key Evaluation Checklist. The formula is simple, the task is not; but it may be better to specify the core of evaluation training in this way rather than by listing competencies in terms of their supposed prerequisite status with respect to evaluation. People get to be good evaluators by a large number of routes, and the field would probably benefit by increasing this number rather than standardizing the routes. See **Evaluation Skills.**

TRAIT-TREATMENT INTERACTION A less widely-used term for *aptitude-treatment interaction*, though it is actually a more accurate term.

TRACERS Artifically added features of a treatment designed to make the identification of its effects easier. See **Modus Operandi Method.**

TRANSACTIONAL EVALUATION (Rippey) Focuses on the process of improvement, e.g. by encouraging anonymous feedback for those that a change would affect, and then a group process to resolve differences. Though a potentially useful *implementation* methodology in some cases, transactional evaluation does not help much with e.g. product evaluation or (in general) with the consumer effects of a program, being mainly staff-oriented.

TRANSCOGNITIVE See **Hypercognitive.**

TREATMENT A term generalized from medical research to cover whatever it is that we're investigating; in particular whatever is being applied or supplied to, or done by, the experimental group that is intended to distinguish them from the comparison group(s). Using a particular brand of toothpaste or toothbrush or reading an advertisement or textbook or going to school are all examples of treatments. "**Evaluand**" covers these, but also products, plans, and people etc.

TRIANGULATION Originally the procedure used by

surveyors to locate ("fix") a point on a grid. In evaluation, or scientific research in general, it refers to the attempt to get a fix on a phenomenon by approaching it from more than one independently based route. For example, if you want to ascertain the extent of sex stereotyping in a company, you will interview at several levels, you will examine training manuals and interoffice memos, you will observe personnel interviews and files, you will analyze job/sex/qualification matches, job descriptions, advertising, placement and so on. In short, you avoid commitment to the validity of any one source by the process of triangulation.

TRUE CONSUMER Someone who, directly or indirectly, receives the services etc. provided by the evaluand. Does not include the service providers though they are also part of the **impacted population**. Is usually a very different group from the target population (intended consumers.)

TRUE EXPERIMENT A "true experiment" or "true experimental design" is one in which the subjects are matched in pairs or by groups as closely as possible and then one from each pair or one group is *randomly assigned* to (be) the control and the other to the experimental group. The looser-and-larger numbers version skips the matching step and just assigns subjects randomly to each group. (Cf. **ex post facto** design and **quasi-experimental** design.

TWO-TIER SYSTEM (also called Multi-Tier System, and Hierarchical System) A system of evaluation, sometimes used in proposal evaluation, (but also with considerable potential in personnel evaluation) where an attempt is made to reduce the total social cost of the ordinary RFP system by requiring two rounds of competition. The first, which is the only one RFP'd, involves stringent length restrictions on the proposal, which is supposed to indicate just the general approach and, e.g., personnel available. These brief sketches are then reviewed by panels that can move through them very fast, and a small number of promising ones are identified. Grants are (sometimes) made to the authors of this "short list" of bidders in order to cover their costs in developing full proposals. The relatively small number of full proposals is then reviewed by a (sometimes considerably smaller) group of reviewers or reviewing panels—the second tier of the review system. The mathematics

of this varies from case to case, but it's worth looking at an example. Suppose we simply put out the usual kind of RFP for improvement of college science teaching laboratories. We might get back 600 or 1,200 proposals, averaging perhaps 50 or 60 pages in length. For convenience let's say they average 50 pages and we get 1,000 of them. That's 50,000 pages of proposals to be read, and 50,000 pages of proposals to be written. Even if reviewers can "read" 100 or 200 pages an hour, we're still looking at 250–500 hours of proposal reading, which means about 60 person-days of reading, i.e. a panel of 15 working for four days, two panels of 15 working two days, or ten panels of 6 working for one day. The problem is that you can't get good reviewers for four days; and the small panels require more personnel to staff, and then have to face the serious problem of interpanel differences. Now if we go to a two-tier system, then we can place an upper limit of, say, five pages on the first proposal and, although we may get a few more, that's a good result since it means that we'll get some entries who don't have the time or resources required to submit massive proposals. So we might start with 1,200 five-pagers, which is 6,000 pages, and we've immediately got a reduction of 88 percent in the amount of reading that's done, with the result that a single panel can reasonably manage it. Then there will be perhaps ten or twenty best proposals coming in at the 50 page length, which can be handled quite quickly, and indeed much more carefully, by the same panel, reconvened for that purpose. Notice also that the reading speed for the first tier of proposals may be higher since all the readers have to do is to be sure they're not missing a promising proposal, rather than to rank-order for final award. And validity should be higher. Notice the triple savings that are involved: the proposers can save about 90 percent of their costs (it may not be quite so high, because shorter proposals take more than a prorated-by-page amount of resources, but it's still substantial); the agency saves a great deal of cost in paying raters or panelists, and heavy staff work costs; and the reliability of the process as well as the quality of available judges goes up significantly. Hence the small subsidy for the second tier proposal is more than justified, both fiscally and in terms of encouraging entries from people that couldn't otherwise afford it; and better entries for those that can.

TYPE 1/TYPE 2 ERRORS See **Hypothesis Testing.**

UNANTICIPATED OUTCOMES Often used as a synonym for **side-effects**, but only loosely equivalent, since: outcomes may be unanticipated by inexperienced planners but readily predictable by experienced ones; effects that are anticipated but not goals are (sometimes) still side-effects—and sometimes not (e.g. having to rent offices.)

UNCERTAINTY, Evaluating. See **Risk.**

UNOBTRUSIVE MEASUREMENT The opposite of **reactive** measurement. One that produces no reactive effect, e.g. observing the relative amount of wear on the carpet in front of interactive displays in a science museum as a measure of relative amounts of use. Sometimes unethical, and sometimes ethically preferable to obtrusive evaluation. ("Obtrusive" is not necessarily "intrusive"; it may be obvious but not disruptive.)

UTILITY (Econ.) The value of something to someone or to some institution. "Interpersonal comparison of utility" is the stumbling-block of (welfare) economics. Sometimes measured in the hypothetical units of "utiles". See - **Apportionment.**

UTILIZATION (of evaluations). This refers *either* to the effort to improve implementation of an evaluation's recommendations *or* to a metaevaluative focus on the extent to which evaluations have been utilized. Utilization/implementation must be planned into evaluations from the first moment; indeed, if the client isn't in a position to utilize the results appropriately, an ethical question arises as to whether the evaluation should be done. Standard procedures include putting representatives of the evaluees on the evaluation team or advisory panel; soliciting and using suggestions from the whole impacted population about design and findings; identifying and focusing on positive benefits of the evaluation if implemented; using appropriate language, length and formats in the report(s); establishing a **balance of power** to reduce threat; and, most importantly, a heavy emphasis on explaining/teaching about the particular and general advantages of evaluation. See also **Implementation of Evaluations.**

VALIDITY A test is valid if it really does measure what it purports to measure. It can be reliable (in the technical sense) without being valid, and it can be valid without being credible. But if it's valid it has to be reliable—if the thermometer is valid, it must say 100 degrees Centigrade *whenever* placed in boiling water and hence must agree with itself, i.e. be reliable. There are various subspecies of validity in the jargon (especially **face, content, construct,** and **predictive** validity,) but they represent an inflation of methodological differences into supposed conceptual distinctions, except perhaps "face-valid" which possibly should be distinguished since it only means "look valid." Serious investigation of validity will identify the appropriate kind for the (e.g.) test being studied; one should not talk about "valid in *this* sense, but not in *that*," only about "*valid* in the appropriate sense."

VALUED PERFORMANCE A value-imbued descriptive variable, imbued with value by the *context*. For example, in the context of evaluating hot rods, the standing-start quarter-mile time is the principal evaluative measure, the valued performance. On the one hand it's totally factual/descriptive; on the other hand, it is contextually imbued with value and is treated exactly as if it logically involved the concept of merit. Cf. **Crypto-evaluative term.**

VALUE-FREE CONCEPTION OF SCIENCE The belief that science, and in particular the social sciences, should not or cannot properly infer to evaluative conclusions, on the basis of purely scientific considerations. Mistakenly assumed to be a consequence of empiricism though in fact it requires the further (erroneous) premise that inference from facts to values is impossible; he error is precisely analogous to the error of supposing that one cannot infer to conclusions about theoretical constructs from observations. (Popper's simplistic attack on induction is thus partly responsible for the continued support of the value-free doctrine.) Apart from the logical errors, there is the evidence of one's senses that science is redolent with highly responsible and well-justified scientific evaluation of research designs, of estimates, of fit, of instruments, of explanations, of research quality, of theories. That the value-free position was maintained at all in the face of these considerations, requires an explanation in terms of **valuephobia.** See **Needs Assess-**

ment and *LE.*

VALUE-IMBUED TERM See **Valued Performance.**

VALUEPHOBIA The resistance to evaluation that generated the myth of value-free science, the attacks on properly-used testing or course grading (see **Kill the Messenger**), on program evaluations for accountability and on the evaluation of college faculty is often more than any rational explanation can cover. We use the term "valuephobia" to cover it without any implications of neurosis, just irrationality. Of course the natural defensive strategy (attack anything that is a threat) is part of it; but part of it goes deeper, into the unwillingness to face possibly unpleasant facts about oneself even if it means large long-run benefits. (This phenomenon—related to "denial"—is seen in people who won't go to a doctor because they don't want to *hear* about imperfections). Valuephobia leads to many abuses e.g. pathetic guarantees that an evaluation will be done "only to help, not to criticize" (if there are no valid criticisms, there's rarely any justification for help of programs/performances involving professionals); to the substitution of implementation monitoring for outcome-based program evaluations; to the refusal of professional associations to use professional standards in their own accreditation or enforcement procedures; to excessive involvement of evaluation staff with the program staff ("to reduce anxiety or "to improve implementation" ") which frequently produces pablum evaluations; & (via guilt) to the absurd ratio of favorable to unfavorable program evaluations—absurd given what we really know about the proportion of bad programs. The clinical status of valuephobia as a U.S. cultural phenomenon is more obvious to a visitor from e.g. England where very tough criticism in the academy is not taken personally to the degree it is here; and it is in this country that Consumers Union was listed by the Attorney-General as a subversive organization and (independently) banned from advertising in newspapers. But the ubiquity of valuephobia is more important; Socrates was killed for his teaching and application of evaluative skills and dictators today seem no less inclined to murder their critics than the Greek "democracy." Humility may best be construed not as the avoidance of self-regard but as the valuing of criticism: the outcome of successful "treatment" (hopefully educational rather than

therapeutic) for valuephobia; this should be combined with some capacity to distinguish good from bad criticism. See **Educational Role**.

VALUES (in evaluation; & measurement of) The values that make evaluations more than mere descriptions can come from a variety of different sources. They may be picked up from a relevant and well-tried set of e.g. professional standards. They may come from a **needs assessment** which might show that children become very ill without a particular dietary component (i.e. need it). Or they may come from a logical and pragmatic analysis of the function of something (processing speed in a computer is a virtue, *ceteris paribus*.) They may even come from a study of wants *and* of the absence of ethical impediments to their fulfilment (e.g. in building a better roller-coaster.) In each of these cases, the foundations are factual and the reasoning is logical—nothing comes in that a scientist should be ashamed of. But something hovers in the background that scientists are embarrassingly incompetent to handle, namely ethics. Without *doing* ethics, however, *most* evaluations can be validated by just checking for salient ethical considerations that might override the non-ethical reasoning. The values/preferences that sometimes come into the evaluation as the ultimate data range in visibility for obvious (political ballots) to very inaccessible (attitudes towards job-security, women supervisors, censorship of pornography.) Most instruments for identifying the more subtle ones are of extremely dubious validity; they are best inferred from behavior; although that inference is also difficult, it begins with the kind of event we are (usually) hoping to influence. Some simulations are so good that they probably elicit true values, especially if not very important ones are involved; usually behavior in real situations should be used.

VARIABILITY The extent to which a population is spread out over its range, as opposed to concentrated near one or a few places (or modes)—the feature that produces **dispersion.**

WHOLISTIC Alternative spelling of **Holistic**.

WHY DENY A conference with the staff of a funding

agency which unsuccessful bidders on an RFP may request and at which they are informed about the reasons why they lost out. One of the consequences of the recent move towards openness. Unfortunately the failure to use salience scoring and other systematic procedures means that reviewer and staff feedback is very difficult to interpret in a useful way.

WIRED A contract or an RFP is said to be "wired" if either through its design and requirements or through an informal agreement between agency staff and a particular contractor, it is arranged so that it will go to that contractor. Certainly illegal, and nearly always immoral. The mere fact that the RFP—with intrinsic good reasons—pre-determines the contractor e.g. because the problem can in fact only be handled by an outfit with two Cray computers, does not constitute wiring.

WORTH System value by contrast with intrinsic value i.e. **merit**; e.g. market value is a function of the market, not of the evaluand's own virtues. The worth of a professor is a function of the enrollment in her or his classes, grant-getting, relation to the college's mission, role-modeling function for prospective/actual women or minority students, *as well as* his/her professional merit. The latter is a necessary but not sufficient condition for the former.

ZERO-BASED BUDGETING (ZBB) A system of budgeting in which *all* expenditures have to be justified rather than *additional* expenditures (i.e. variations from "level-funding".)

ACRONYMS & ABBREVIATIONS

AA Audit Agency—a division of HHS that reports directly to the Secretary and does internal audits (cf. GAO) that amount to evaluations of program efforts and contracts including evaluations. Has moved from CPA orientation to much broader approach and does much very competent work (though spread a little thin); still doesn't look at e.g. validity of test-instruments used.

AAHE American Association of Higher Education

ABT Properly, Abt Associates. Large shop with strong evaluation capability; headquarters Cambridge, Massachusetts.

ACT American College Testing—big Iowa-based educational measurement shop.

AERA American Educational Research Association

AID Agency for International Development

AIR American Institutes for Research, a Northern California-based contractor with some evaluation capability.

ANCOVA Analysis of covariance

ANOVA Analysis of variance

ATI Aptitude-treatment interaction

AV Audiovisual

AVLINE Online audio-visual database maintained by

NLM

CAI Computer-assisted instruction

CBO Congressional Budget Office. Provides analysis and evaluation services to Congress, as GAO does for the administration.

CBTE, CBTT, CBTP Competency Based Teacher Education, Training or Preparation

CDC Computer Development Corporation; one of the top five computer companies.

CEEB College Entrance Examination Board

CEDR Center for Evaluation, Development and Research (at Phi Delta Kappa)

CFE Cost-free evaluation

CIPP Daniel Stufflebeam and Egon Guba's model which distinguished four types of evaluation: context, input, process, and product—all designed to delineate, obtain, and provide useful information for the decision-maker.

CIRCE Center for Instructional Research and Curriculum Evaluation, University of Illinois, Urbana, Illinois.

CMHC Community Mental Health Center or Clinic

CN ***Consultants News***, the highly independent newsletter of the management consulting area, run by talented loner Jim Kennedy.

COB Close of business (end of working day; a proposal deadline)

CRT Criterion-referenced test (or cathode ray tube, the display monitor on some computers)

CSE Center for the Study of Evaluation (at UCLA)

CSMP Comprehensive School Mathematics Study Group

DEd (properly ED) Department of Education (ex-USOE)

DOD Department of Defense

DOE Department of Energy

DRG Division of Research Grants

DRT Domain-referenced test

ED Education Department

EIR Environmental Impact Report

EN ***Evaluation News***, the newsletter of the Evaluation Network

ENet Evaluation Network, an organization of evaluators

EPIE Education Products Information Exchange

ERIC Educational Resources Information Center; a nationwide information network with its base in Washington, D.C. and 16 clearinghouses at various locations in the U.S.

ERS Evaluation Research Society

ESEA Elementary and Secondary Education Act of 1965

ETS Educational Testing Service; headquarters in Princeton, N.J.—branches in Berkeley, Atlanta, etc.

FRACHE Federation of Regional Accrediting Commissions of Higher Education

FY Fiscal year

G & A General and administration (expenses, costs)

GAO General Accounting Office. The principal semi-external evaluation agency of the Federal government.

GBE Goal-based evaluation

GFE Goal-free evaluation

GIGO Garbage In, Garbage Out (from computer programming; see **meta-analysis**)

GPA Grade-point average

GPO Government Printing Office, Washington, D.C.

GRE Graduate Record Examination

HEW Department of Health, Education and Welfare, now divided into E.D. and H.H.S.

HHS Department of Health and Human Services

IBM International Business Machines.

IOX Instructional Objectives Exchange in Los Angeles

K $1000 as in "16K for evaluation."

K-12 Kindergarten through high school years

K-6 The domain of elementary education

LE ***The Logic of Evaluation***, a monograph by the present author in this series

LEA Local Education Authority (e.g. school district)

Law School Admission Test

M Thousand, as in"$16M for evaluation."

MAS Management Advisory Services; term usually refers to subsidiaries of the Big 8 accounting firms.

MBO Management by Objectives

MCT Minimum Competency Testing

MIS Management Information System; usually a computerized database combining fiscal, inventory, and performance data.

MMPI Minnesota Multiphasic Personality Inventory

MOM Modus Operandi Method

NCES National Center for Educational Statistics

NCHCT National Center for Health Care Technology

NIA National Institute onAging

NICHD National Institute of Child, Health and Human Development

NIE National Institute of Education (in ED)

NIH National Institutes of Health (includes NIMH, NIA etc.), *or* Not Invented Here (so don't encourage its use because someone else will get the credit and people will think we can't manage our own affairs.)

NIJ National Institute of Justice

NIMH National Institute of Mental Health

NLM National Library of Medicine

NSF National Science Foundation

NWL Northwest Lab, Portland, Oregon. One of the federal network of labs and R & D centers; currently has strongest evaluation staff.

OE Office of Education

OHDS Office of Human Development Services

OJT On-job training

OMB Office of Management and Budget

OPB Office of Planning and Budgeting

ONR Office of Naval Research; sponsor of e.g. ***Encyclopedia of Educational Evaluation.***

OTA Office of Technology Assessment

P & E Planning and Evaluation; a division of HEW/HHS, including regional offices, where it reports directly to Regional Directors. In ED, currently called OPB

PBTE Performance Based Teacher Education

PDK Phi Delta Kappa, the influential and quality-oriented educational honorary.

PEC Product Evaluation Checklist, forerunner of KEC.

PERT Program Evaluation and Review Technique

PHS Public Health Service

PLATO The largest CAI project ever; headquarters at the University of Illinois/Champaign. Mostly NSF funded in development phase, now CDC-controlled.

PPBS Planning-Programming-Budgeting-System

PSI Personalized System of Instruction (a.k.a. The Keller Plan)

PT Programmed Text

RAND Big Santa Monica-based contract research and evaluation and policy analysis outfit. Originally, a U.S. Air-Force 'creature' (civilian subsidiary), set up because they couldn't get enough specialized talent from within the ranks—name came from **R**esearch **A**nd **D**evelopment. Now independent non-profit, though still does some work for USAF.

RFP Request for proposal

SAT Scholastic Aptitude Test. Widely used for college admissions.

SDC Systems Development Corporation in Santa Monica; another large shop like Rand with substantial evaluation capability.

SEA State Education Authority

SEP School Evaluation Profile

SES Socioeconomic status

SMSG School Mathematics Study Group. One of the earliest and most prolific of the federal curriculum reform efforts.

SRI Originally Stanford Research Institute; in Menlo Park, CA; once part-owned by Stanford University, now autonomous. Large "shop" which does some evaluation.

TA Technology Assessment *or* Technical Assistance *or* Teaching Assistant

TAT Thematic Apperception Test

TCITY Twin Cities Institute for Talented Youth. Site of the first advocate-adversary evaluation.

USAF United States Air Force. Heavy R & D commitment, like Navy, and unlike Army or Marine Corps

USDA United States Department of Agriculture

USOE United States Office of Education, now ED or DEd (Department of Education)